On Path to Multimodal Generalist: General-Level and General-Bench

Hao Fei [*1] **Yuan Zhou** [*2] **Juncheng Li** [*3] **Xiangtai Li** [*2] **Qingshan Xu** [*2] **Bobo Li** [*1] **Shengqiong Wu** [*1] **Yaoting Wang** [4] **Junbao Zhou** [2] **Jiahao Meng** [5] **Qingyu Shi** [5] **Zhiyuan Zhou** [6] **Liangtao Shi** [6] **Minghe Gao** [3] **Daoan Zhang** [7] **Zhiqi Ge** [3] **Siliang Tang** [3] **Kaihang Pan** [3] **Yaobo Ye** [3] **Haobo Yuan** [2] **Tao Zhang** [8] **Weiming Wu** [9] **Tianjie Ju** [10] **Zixiang Meng** [8] **Shilin Xu** [5] **Liyu Jia** [2] **Wentao Hu** [2] **Meng Luo** [1] **Jiebo Luo** [7] **Tat-Seng Chua** [1] **Shuicheng Yan** [1 11] **Hanwang Zhang** [2]

Abstract

The Multimodal Large Language Model (MLLM) is currently experiencing rapid growth, driven by the advanced capabilities of language-based LLMs. Unlike their specialist predecessors, existing MLLMs are evolving towards a Multimodal Generalist paradigm. Initially limited to understanding multiple modalities, these models have advanced to not only comprehend but also generate across modalities. Their capabilities have expanded from coarse-grained to fine-grained multimodal understanding and from supporting singular modalities to accommodating a wide array of or even arbitrary modalities. To assess the capabilities of various MLLMs, a diverse array of benchmark test sets has been proposed. This leads to a critical question: *Can we simply assume that higher performance across tasks indicates a stronger MLLM capability, bringing us closer to human-level AI?*

We argue that the answer is not as straightforward as it seems. In this project, we introduce an evaluation framework to delineate the capabilities and behaviors of current multimodal generalists. This framework, named **General-Level**, establishes 5-scale levels of MLLM performance and generality, offering a methodology to compare MLLMs and gauge the progress of existing systems towards more robust multimodal generalists and, ultimately, towards AGI (Artificial General Intelligence). Central to our framework is the use of **Synergy** as the evaluative criterion, categorizing capabilities based on whether MLLMs preserve synergy across comprehension and generation, as well as across multimodal interactions. To evaluate the comprehensive abilities of various generalists, we present a massive multimodal benchmark, **General-Bench**, which encompasses a broader spectrum of skills, modalities, formats, and capabilities, including over 700 tasks and 325,800 instances. The evaluation results that involve over 100 existing state-of-the-art MLLMs uncover the capability rankings of generalists, highlighting the challenges in reaching genuine AI. We expect this project to pave the way for future research on next-generation multimodal foundation models, providing a robust infrastructure to accelerate the realization of AGI.

Project Page: https://generalist.top/

Leaderboard: https://generalist.top/leaderboard/

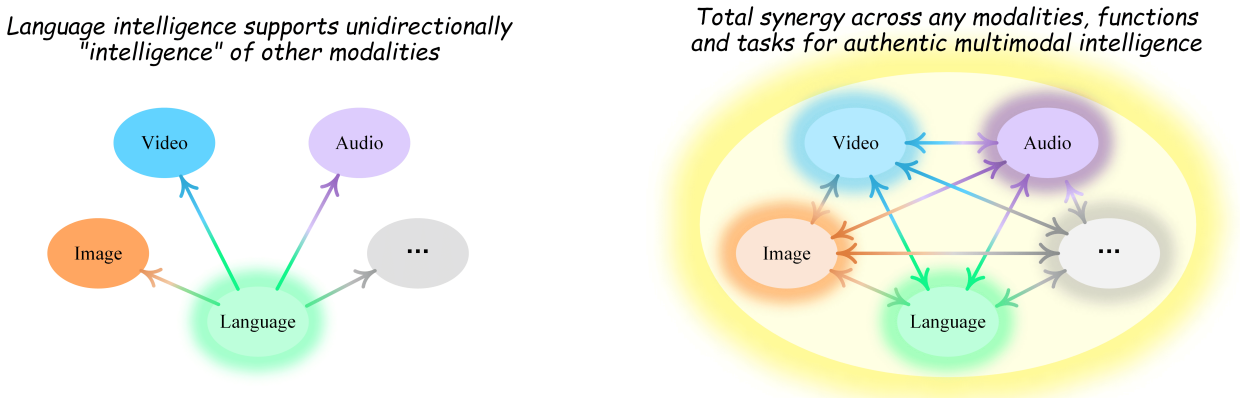
Benchmark: https://huggingface.co/General-Level/

1 Introduction

Large Language Models (LLMs, e.g., ChatGPT (OpenAI, 2022a) and LLaMA (Touvron et al., 2023)) have revolutionized the NLP field by serving as generalists addressing a vast spectrum of NLP tasks. This breadth of capability has edged humans ever closer to the realization of Artificial General Intelligence (AGI). Yet, human intelligence inherently operates across multiple modalities, not solely through language. This observation has spurred the development of multimodal LLMs (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023a; OpenAI, 2022b), i.e., multimodal generalists, which are rapidly gaining traction and evolving towards AGI. The recent progress in MLLMs is marked by significant advancements. For example, the initial multimodal agents where LLMs serve as mere task schedulers, later have evolved into joint foundation MLLMs (Zhu et al., 2023a; Liu et al., 2023a; Zhang et al., 2023a; OpenAI, 2022b; Wu et al., 2024a; Chen et al., 2024a; Sun et al., 2024). Also,

[*]Equal contribution [1]NUS [2]NTU [3]ZJU [4]KAUST [5]PKU [6]HFUT [7]UR [8]WHU [9]NJU [10]SJTU [11]Skywork AI. Correspondence to: Shuicheng Yan <yansc@nus.edu.sg>, Hanwang Zhang <hanwangzhang@ntu.edu.sg>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.



(*a*) Existing intelligent pattern in multimodal generalist

(*b*) Ideal intelligent pattern in multimodal generalist

Figure 1: The "*intelligence*" in most existing multimodal generalists (i.e., MLLMs) hinges on language intelligence (i.e., from LLMs) **(a)**, whereas the ideal intelligence mode should be maintaining synergy across all modalities and tasks **(b)**.

MLLMs have progressed from understanding only multimodal signals to both comprehending and generating multimodal content, even editing capabilities (Wang et al., 2023a; Munasinghe et al., 2023; Zhang et al., 2024a; Fei et al., 2024a). Further, these models have advanced from coarse-grained modal understanding to fine-grained multimodal comprehension, such as pixel-level visual modeling (Ren et al., 2023; Yuan et al., 2023a; Rasheed et al., 2023). More significantly, MLLMs that initially support only singleton non-textual modalities have now facilitated the understanding and generation of signals across various modalities, even simultaneously accommodating any modality (Wu et al., 2024a; Zhan et al., 2024; Lu et al., 2024a).

Accordingly, the community has introduced various benchmarks to evaluate those MLLMs (Wu et al., 2023a; Xia et al., 2024a; Yue et al., 2024; Meng et al., 2024a; Liu et al., 2025; Li et al., 2024a; Ying et al., 2024a; Li et al., 2024b). The prevailing evaluation mindset might yet be largely outdated, simplistically assuming that superior performance across tasks presents a stronger generalist capability (Xu et al., 2023a; Yu et al., 2023; Fu et al., 2024a; Chen et al., 2024b), and then being closer to AGI. We contend this perspective overly simplifies the implication inherent in real multimodal generalization. Theoretically, it's effortless to assemble a "super agent" from all singleton state-of-the-art (SoTA) specialists to achieve the above goal, while such a simplistic integration would never suffice to realize genuine AGI. We argue that the key to advancing towards AGI lies in the ***synergy*** effect—a capability that enables knowledge learned in one modality or task to generalize and enhance mastery in other modalities or tasks, fostering mutual improvement across different modalities and tasks through interconnected learning.[1] As illustrated in Figure 1, most current MLLMs predominantly build on the language intelligence of LLMs to simulate the indirect intelligence of multimodality, which is merely extending language intelligence to aid multimodal understanding. While LLMs (e.g., ChatGPT) have already demonstrated such synergy in NLP, reflecting language intelligence, unfortunately, the vast majority of MLLMs do not really achieve it across modalities and tasks.

In this project, we introduce a sophisticated evaluation framework, **`General-Level`**, for more accurately positioning and assessing the capabilities of current MLLM generalists, charting a path toward authentic multimodal AGI. Drawing inspiration from the tiered classification mechanism in the automotive industry for autonomous vehicles (Yurtsever et al., 2020), `General-Level` defines five principal levels of model performance and generality. Central to the framework is the synergy ability as the evaluative criterion, categorizing capabilities based on whether generalists preserve synergy in and across multimodal comprehension and generation, as well as cross-modal interactions. From the lowest to the highest level, the scope of synergy ability required progressively escalates from single tasks or modalities to total synergy. As a generalist strives to advance to a higher level, it must demonstrate significant enhancements in its synergy capabilities, during which the difficulty of progression is also inherently increasing.

To effectively evaluate within the `General-Level` framework, a suitable benchmark is essential. While there are numerous MLLM evaluation benchmarks, e.g., LVLM-eHub (Xu et al., 2023a), MME (Fu et al., 2024a), MMMU (Yue et al., 2024), SEED-Bench (Li et al., 2024a), MMT-Bench (Ying et al., 2024a), and MEGA-Bench (Chen et al., 2024b), they might have certain limitations that render them inadequate for our needs. Firstly, existing benchmarks often convert all tasks into a uniform multiple-choice QA format (Fu et al., 2024a; Ying et al., 2024a), simplifying the evalua-

[1] Synergy, in essence, can be understood as a form of generalization ability.

tion process but consequently restricting assessments to only the models' multimodal comprehension capabilities. However, a true multimodal generalist should support not only comprehension, but also possess capabilities in multimodal generation, editing, and beyond. Second, the majority of current benchmarks (Wu et al., 2023a; Liu et al., 2025; Li et al., 2024a) predominantly focus on the image modality and overlook other crucial modalities such as video, audio, even 3D and beyond, which are vital for a robust multimodal generalist. Third, these benchmarks are typically limited to coarse-grained multimodal understanding (Xu et al., 2023a; Yu et al., 2023; Fu et al., 2024a) and fail to adequately assess finer-grained ones, which actually lag far behind the current advancements in MLLMs, i.e., supporting pixel-level image understanding and generation (Fei et al., 2024a; Zhang et al., 2024a). In response to these challenges, we propose **`General-Bench`**, a massive multimodal evaluation benchmark, spanning from various modalities (e.g., image, video, audio, 3D, language, and beyond) in diverse native formats, covering a wide range of tasks that thoroughly assess the full capabilities of a multimodal generalist.

Our evaluation of over 100 existing top-performing LLM/MLLM systems has uncovered critical insights into their capabilities and rankings as multimodal generalists. The most notable finding is that most MLLMs lack the cross-task or cross-modal synergy ability required for higher-level classifications, with even advanced models like GPT-4V and GPT-4o not achieving top ranks. This highlights a considerable gap in achieving the goals of multimodal generalists. Also, the majority of existing MLLMs manage only a few basic multimodal tasks and skills, which negatively affects their scoring. Most critically, no model has yet demonstrated the ability to enhance language intelligence through non-language modalities, underscoring the substantial challenges in the pursuit of genuine AGI.

Contributions: 1) We introduce a tiered classification system called `General-Level` for multimodal generalists, establishing a rigorous norm that can guide future MLLM research. 2) We contribute a new evaluation benchmark (`General-Bench`) that provides the most comprehensive coverage of modalities and tasks available to date.

2 Background and Related Work

More and more tend to recognize that LLMs have unlocked the potential of language intelligence, bringing unprecedented hope to achieve AGI. Essentially, an LLM serves as a generalist capable of tackling nearly all downstream NLP tasks. LLMs have subsequently evolved in an effort to extend this intelligence across various other modalities, i.e., MLLMs (Bai et al., 2023; Zhang et al., 2023b; Jin et al., 2023; Li et al., 2024c; Fei et al., 2024b;c). Unlike the past 'smaller' specialists (Van Den Oord et al., 2016; Radford et al., 2021; Rombach et al., 2022; Liu et al., 2023b), MLLMs represent an important advancement of unification to handle all modalities and tasks with one foundation model, i.e., multimodal generalists. Naturally, empowering a multimodal generalist with strong multimodal intelligence capabilities is an essential pathway toward realizing AGI.

Technically, the vast majority of existing MLLMs have frameworks that are anchored by an LLM to serve as the core for reasoning and decision-making. By integrating various well-trained modules of different modalities or tasks (typically existing specialists, e.g., CLIP (Radford et al., 2021) and Stable Diffusion (Rombach et al., 2022)), MLLMs are facilitated with the comprehension and even generation of diverse modalities. Representative MLLMs include Blip2 (Li et al., 2023a), LLAVA (Liu et al., 2023a), MiniGPT-4 (Zhu et al., 2023a), Flamingo (Alayrac et al., 2022), and NExT-GPT (Wu et al., 2024a), among others. However, such an architectural setup merely simulates 'pseudo' multimodal intelligence, as it still fundamentally relies on the language intelligence of LLMs without genuine non-language modality intelligence. As emphasized earlier, a capable generalist must possess synergy capabilities across all modalities and tasks, akin to how an LLM (e.g., ChatGPT) generalizes well to unseen NLP tasks, despite not being exposed to all tasks during its training. While these current multimodal generalists can deliver strong performances on multimodal benchmarks, sometimes even on par with SoTA specialists, they do not fundamentally achieve true synergy.

Consequently, this paper positions synergy as the central criterion for evaluating multimodal generalists on their journey toward AGI. Current evaluation methods (Li et al., 2024b) for MLLMs still adhere to the traditional approach used for specialists, simply comparing the MLLM performance on multimodal tasks, assuming that higher scores indicate greater strength and closer proximity to AGI. Going beyond that, we propose a new evaluation framework—not only do we compare whether models support various modalities and tasks and their performance, but we also rank them based on the synergy capabilities of multimodal generalists. Meanwhile, we significantly expand the scope of current MLLM benchmark datasets in terms of modality and task coverages, as well as task formats, contributing to the most comprehensive benchmark dataset to date in the community.

3 *General-Level*: A 5-Level Taxonomy of Multimodal Generalists

3.1 Preliminary

3.1.1 Observations and Principles

Observation-1: Multimodal Comprehension vs. Simultaneous Multimodal Comprehension and Generation. Initially, MLLMs are capable only of interpreting multimodal signals, meaning their responses are limited to textual

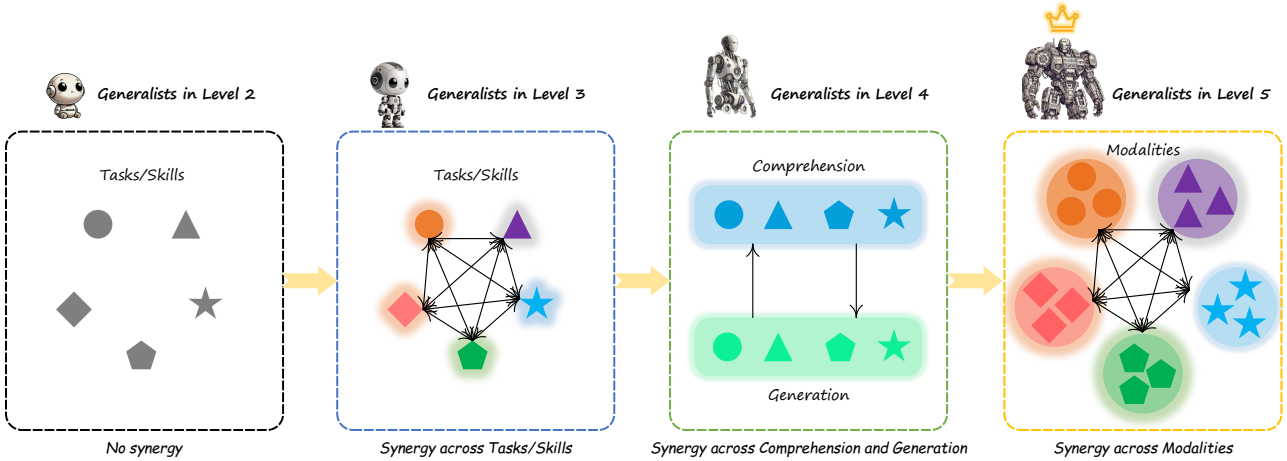


Figure 2: A specific illustration on **synergy** effect.

outputs based on user-provided multimodal inputs. However, an MLLM that only offers multimodal comprehension operates at the most basic and rudimentary level. More advanced MLLMs have since emerged, equipped with not only multimodal comprehension but also the ability to generate and even edit content across various modalities. It is widely believed that the more advanced a multimodal generalist is, the more it should encompass advanced functionalities, encompassing both comprehension and generation.

Observation-2: Covering Broader Modalities. Being a multimodal generalist requires the ability to extensively support and handle a wide range of modal data, including, but not limited to, text, images, videos, audio, and even 3D. The extent of modal support is indicative of the breadth of an AI system's capabilities. Initially, MLLMs could manage only a singleton non-linguistic modality, e.g., images, videos, or audio signals. To date, these models have evolved to simultaneously support multiple non-linguistic modalities—such as combining images with videos, videos with audio, and even any modality in the current most advanced cases.

Observation-3: Supporting Various Tasks and Paradigms. To qualify as a true multimodal generalist, it must be capable of handling a broad range of tasks with different definitions and requirements. The greater the variety of tasks supported, the stronger the generalist's overall versatility. For example, early visual MLLMs could only manage coarse-grained image understanding, but recent advancements have enabled them to achieve fine-grained, pixel-level multimodal comprehension, such as pixel-level image/video grounding and editing. This advancement necessitates that the model's decoding components should be versatile enough to generate outputs in various task formats, not merely restricted to text.

Observation-4: Multimodal Agent vs. Multimodal Foundation Model. Initially, researchers approach multimodal tasks by using LLMs as task schedulers, where an LLM orchestrates the execution of tasks by invoking external tools and modules (often specialists) to handle specific multimodal tasks. This setup is referred to as a multimodal agent. Subsequently, attention shifted towards building joint MLLMs, where the LLM is tightly integrated with other modules, such as multimodal understanding components (front-end) and multimodal generation components (back-end), through a shared embedding space. This setup allows for joint training, where the entire system, including all parameters, can be updated end-to-end. The complexity of AGI requires deeper integration and generalization across tasks and modalities.

3.1.2 Synergy as Core to Generalists

We argue that determining whether a multimodal generalist is stronger cannot be simplistically equated with achieving higher scores on a benchmark or/and supporting as many multimodal tasks as possible compared to other models—a common practice in current MLLM benchmarking and evaluation. A simple counterexample can illustrate this point: it could be comparatively easier to construct a 'super agent' by integrating all SoTA specialists for various multimodal tasks into a single system. Such an agent could achieve top-level performance across all tasks (on par with the strongest individual specialist models) while supporting a wide range of multimodal functionalities. However, such agents can be far from the multimodal generalist we expect as a pathway to AGI. Such a type of agent lacks inherent multimodal intelligence and capabilities, as it relies on an ensemble of specialized systems rather than embodying true, native multimodal generalization.

Instead, the ideal multimodal generalist (and ultimately AGI) we envision should be a multimodal counterpart of an all-capable OpenAI ChatGPT series. Such a model would not only surpass SoTA specialists in task-wise performance across various tasks and modalities but also exhibit exceptional *cross-task*, *cross-comprehension-generation*, and *cross-modality* generalization capabilities. In other words,

Table 1: **`General-Level`** framework toward classifying multimodal generalists into **FIVE** levels based on the synergy abilities models preserve. We denote the number of tasks within the Comprehension group by M; the number within the Generation group by N; and the number of NLP tasks by T.

Level	Definition	Scoring	Example
Level-1: Specialists	Various current models, each fine-tuned on a specific task or dataset of specific modalities, are task-specific players (i.e., SoTA specialists). This includes various learning tasks, such as linguistic/visual recognition, classification, generation, segmentation, grounding, inpainting, and more.	For each task in the benchmark (i-th task), the current SoTA specialist's score is recorded as: σ_i^{sota}	CLIP (Li et al., 2022), FLUX (Labs, 2023), FastSpeech2 (Ren et al., 2021), ···
↓ ***Upgrading Condition***: *Supporting as many tasks and functionalities as possible*			
Level-2: Generalists of Unified Comprehension and/or Generation	Models are task-unified players, e.g., MLLMs, capable of supporting different modalities and tasks. Such MLLMs can integrate various models through existing encoding and decoding technologies to achieve aggregation and unification of various modalities and tasks (such as comprehension and generation tasks).	The average score between Comprehension and Generation tasks (i.e., across all tasks) represents the score at this level. A model that can score non-zero on the data is considered capable of supporting that task. The more supported tasks and the higher the scores, the higher its overall score: $S_2 = \frac{1}{2}\left(\frac{1}{M}\sum_{i=1}^{M}\sigma_i^C + \frac{1}{N}\sum_{j=1}^{N}\sigma_i^G\right)$	Unified-io-2 (Lu et al., 2024a), AnyGPT (Zhan et al., 2024), NExT-GPT (Wu et al., 2024a), SEED-LLaMA (Ge et al., 2023), GPT-4V (OpenAI, 2022b), ···
↓ ***Upgrading Condition***: *Generalists achieving as stronger synergy and cross as many tasks as possible*			
Level-3: Generalists with **synergy** in Comprehension and/or Generation	Models are task-unified players, and synergy is in Comprehension and/or Generation. MLLMs enhance several tasks' performance beyond corresponding SoTA scores through joint learning across multiple tasks due to the synergy effect.	Assign a mask weight of 0 or 1 to each task; mask=1 only if the corresponding score (σ_i^C or σ_i^G) exceeds the SoTA specialist's score, otherwise mask=0. Then, calculate the average score between S_C and S_G. The more tasks to surpass the SoTA specialist, the higher the S_3: $S_3 = \frac{1}{2}(S_G + S_C)$, where $S_C = \frac{1}{M}\sum_{i=1}^{M}\begin{cases}\sigma_i^C & \text{if } \sigma_i^C \geq \sigma_{sota}^C \\ 0 & \text{otherwise}\end{cases}$ $S_G = \frac{1}{N}\sum_{j=1}^{N}\begin{cases}\sigma_j^G & \text{if } \sigma_j^G \geq \sigma_{sota}^G \\ 0 & \text{otherwise}\end{cases}$	GPT-4o (OpenAI, 2022b), Gemini-1.5 (Team et al., 2024a), Claude-3.5 (Team, 2024), DeepSeek-VL (Lu et al., 2024b), LLaVA-One-Vision (Li et al., 2024d), Qwen2-VL (Wang et al., 2024a), InternVL2.5 (Chen et al., 2024c), Phi-3.5-Vision (Abdin et al., 2024), ···
↓ ***Upgrading Condition***: *Generalists in unified comprehension and generation capability with synergy in between*			
Level-4: Generalists with **synergy** across Comprehension and Generation	Models are task-unified players, and synergy is across Comprehension and Generation.	Calculate the harmonic mean between Comprehension and Generation scores. The stronger synergy a model has between Comprehension and Generation tasks, the higher the score: $S_4 = \frac{2S_C S_G}{S_C + S_G}$	Mini-Gemini (Li et al., 2024c), Vitron-V1 (Fei et al., 2024a), Emu2-37B (Sun et al., 2024), ···
↓ ***Upgrading Condition***: *Generalists achieving cross-modal synergy with abductive reasoning ability*			
Level-5: Generalists with **total synergy** across Comprehension, Generation and Language	Models are task-unified players, preserving the synergy effect across Comprehension, Generation, and Language. In other words, the model not only achieves cross-modality synergy between Comprehension and Generation groups but also further realizes synergy with language. The Language intelligence can enhance multimodal intelligence and vice versa; understanding multimodal information can also aid in understanding language.	Calculate the model's average score exceeding SoTA NLP specialists on NLP benchmark data; normalize it to a [0,1] weight, and multiply it by the score from level-4 as the level-5 score: $S_5 = S_4 \times w_L$, where $w_L = \frac{S_L}{S_{\text{total}}}$, where $S_L = \frac{1}{T}\sum_{k=1}^{T}\begin{cases}\sigma_k & \text{if } \sigma_k \geq \sigma_{\text{sota}} \\ 0 & \text{otherwise}\end{cases}$	*None found yet (Let's wait for multimodal ChatGPT moment!)*

the knowledge learned from certain tasks, skills, and modalities should be transferable to other tasks, skills, and modalities—extrapolating the understanding to effectively engage with other tasks and modalities, and vise versa, creating a synergistic effect where the combined result exceeds the sum of individual contributions, achieving a 1+1>2 effect. ChatGPT on the language side can be a good example: it outperforms SoTA specialists in unseen tasks without having undergone specific training for those tasks. This generalizability is what we claim as the **synergy effect**.

3.2 Defining Levels Centered on Synergy

Based on the above principles, we introduce a 5-level taxonomy of multimodal generalists, `General-Level`.

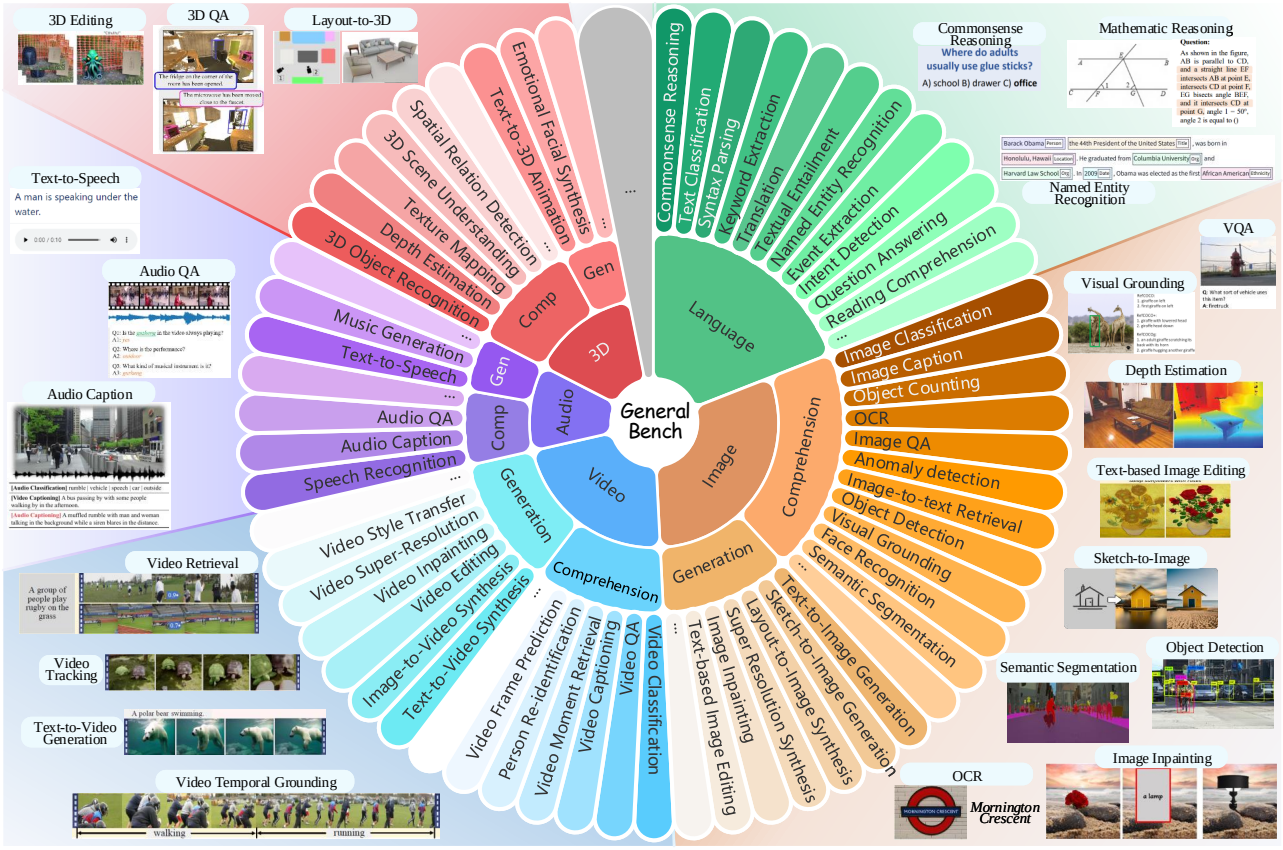


Figure 3: Overview of **General-Bench**, which covers 145 skills for more than 700 tasks with over 325,800 samples under comprehension and generation categories in various modalities. Appendix § B.6 gives holistic hierarchical taxonomies.

General-Level framework evaluates generalists based on the levels and strengths of the synergy they preserve. Specifically, we define three levels and scopes of synergy, ranked from low to high: 'task-task', 'comprehension-generation', and 'modality-modality', as illustrated in Figure 2. Achieving these levels of synergy becomes progressively more challenging, corresponding to higher degrees of general intelligence. Assume we have a benchmark of various modalities and tasks, where we can categorize tasks under these modalities into the Comprehension group and the Generation group, as well as the language (i.e., NLP) group, as illustrated in Figure 5. Now, we can define the scoring specification of `General-Level` as in Table 1.

When calculating scores using the corresponding formula, we normalize all task metrics to a 100-point scale. While most task evaluation scores typically range from 0-100, such as *F1* and *Accuracy*, certain metrics, e.g., *FID*, *MAE*, and *PSNR*, yet yield scores outside this usual range. Thus, we design some mapping functions to standardize performance scores. Our framework also incorporates the principle of diminishing scores: an MLLM (i.e., multimodal generalist) can achieve scores at multiple levels, but it is classified at its highest level, where it achieves a non-zero score.

We assume that current MLLMs have already demonstrated synergy mode from language to non-language modalities. Then the remaining mission is to confirm the existence of synergy in the reverse direction, from non-language to language modalities. Therefore, for level 5—measuring total synergy—we do not measure the generality across all modalities and tasks. Instead, we assess whether a model can improve NLP task performance to exceed that of NLP SoTA specialists.

4 *General-Bench*: A Holistic Benchmark for Multimodal Generalists

We introduce `General-Bench`, a new benchmark to meet the outlined criteria and serve as the standard dataset for our evaluation framework. Appendix §B shows more details about the benchmark data.

The data compiled for `General-Bench` is visualized in Figure 3 visualizes the `General-Bench` highlights of task/modality support. Overall, the current version of the dataset includes those most common modalities (inner ring), and except for NLP tasks, all modalities distinguish between comprehension and generation tasks (middle ring). `General-Bench` particularly places a strong emphasis

Table 2: Performance of multimodal generalists on various image comprehension skills. Skill full names and specific tasks are listed in Appendix. The full performance records of more generalists are shown in Appendix § C.

Model	Image Comprehension Skill (Avg within each #I-C Group)										Task Completion		Level Score on Image		
	#1 #11 #21 #31	#2 #12 #22 #32	#3 #13 #23 #33	#4 #14 #24 #34	#5 #15 #25 #35	#6 #16 #26 #36	#7 #17 #27 #37	#8 #18 #28 #38	#9 #19 #29 #39	#10 #20 #30 #40	#Supported Task	#Win-over-Specialist	Level-2	Level-3	Level-4
SoTA Specialist	51.27 36.40 70.00 39.80	53.32 65.15 50.40 57.20	42.04 43.78 65.97 54.60	22.30 58.90 16.60 63.27	39.02 63.73 78.00 29.60	22.42 87.84 50.48 87.10	46.02 58.66 19.90 98.00	15.67 72.25 53.55 39.60	51.20 34.51 64.10 36.42	28.01 95.70 35.90 82.02	/	/	/	/	/
GPT-4V	69.42 0.00 71.90 40.05	58.64 0.00 37.12 0.00	39.54 51.04 50.30 90.40	0.00 63.52 16.06 0.00	66.18 0.00 72.20 31.64	36.08 70.90 0.00 89.10	61.74 51.60 0.00 22.22	0.00 0.00 72.51 22.54	16.90 0.00 0.00 18.08	20.88 0.00 97.98 84.84	177 (65.1%)	105 (38.6%)	18.16	12.85	0.00
GPT-4o	73.87 0.00 81.30 44.30	63.42 0.00 39.61 0.00	43.23 71.23 48.63 90.40	0.00 61.54 15.12 0.00	71.56 0.00 93.00 33.47	39.65 79.38 0.00 91.20	68.83 55.25 0.00 35.56	0.00 0.00 77.53 24.80	67.80 0.00 0.00 21.12	23.24 0.00 98.79 87.88	177 (65.1%)	112 (41.2%)	19.67	14.51	0.00
Gemini-1.5-Pro	72.33 0.00 84.57 36.41	23.41 0.00 31.55 0.00	39.39 60.86 60.87 98.00	0.00 40.10 15.20 0.00	62.38 0.00 86.40 38.45	34.30 0.00 0.00 92.00	66.25 58.09 0.00 30.37	0.00 0.00 76.72 22.18	59.20 0.00 0.00 21.20	23.79 0.00 96.76 83.23	177 (65.1%)	101 (37.1%)	19.67	12.66	0.00
Gemini-1.5-Flash	67.00 0.00 80.63 28.53	25.79 0.00 28.97 0.00	37.85 55.22 56.91 96.40	0.00 32.92 16.57 0.00	59.45 0.00 82.60 29.97	29.91 0.00 0.00 90.20	63.61 54.57 0.00 27.96	0.00 0.00 73.57 20.64	56.50 0.00 0.00 18.22	22.19 0.00 93.42 80.40	177 (65.1%)	94 (34.6%)	18.54	10.85	0.00
Claude-3.5-Opus	65.38 0.00 70.39 38.28	57.69 0.00 41.19 0.00	39.95 60.21 54.75 91.38	0.00 58.15 13.87 0.00	63.35 0.00 77.80 0.00	34.50 66.57 0.00 87.31	63.43 51.23 0.00 23.87	0.00 0.00 73.04 28.71	45.62 0.00 0.00 25.75	20.44 0.00 94.65 84.65	178 (65.4%)	93 (34.2%)	19.00	11.08	0.00
Emu2-32B	53.76 0.00 56.33 17.73	7.31 0.00 29.43 0.00	36.62 39.47 45.46 72.80	0.00 12.20 21.45 0.00	41.31 0.00 64.20 0.00	22.22 0.00 0.00 73.40	41.89 44.51 0.00 31.72	0.00 5.28 54.59 14.09	21.20 0.00 0.00 18.73	12.83 0.00 70.34 56.97	178 (65.4%)	52 (19.1%)	30.90	5.18	1.25
Phi-3.5-Vision-Instruct	55.32 0.00 67.56 19.31	3.44 0.00 32.32 0.00	34.16 41.00 51.51 83.40	0.00 21.77 23.70 0.00	42.61 0.00 90.10 15.02	42.04 0.00 0.00 80.00	51.34 52.13 0.00 3.98	0.00 11.89 57.68 23.06	0.00 0.00 0.00 25.41	24.35 0.00 52.02 71.31	179 (65.8%)	85 (31.3%)	16.46	9.39	0.00
Qwen2-VL-72B	66.98 0.00 81.86 4.33	5.74 0.00 38.59 0.00	35.64 45.66 58.99 77.64	0.00 29.44 16.17 0.00	56.58 0.00 97.43 16.83	40.50 0.00 0.00 79.34	48.79 59.87 0.00 11.65	0.00 10.89 72.47 29.62	43.18 0.00 0.00 32.22	25.32 0.00 92.41 62.83	177 (65.1%)	99 (36.4%)	19.41	12.34	0.00
SEED-LLaMA-13B	46.68 0.00 38.53 0.00	0.00 0.00 22.52 0.00	31.85 25.42 32.67 71.60	0.00 8.00 24.96 0.00	40.59 0.00 32.20 0.80	13.48 0.00 0.00 69.80	35.10 33.60 0.00 14.43	0.00 4.76 47.48 13.13	7.20 0.00 0.00 10.19	9.09 0.00 66.19 51.72	174 (64.0%)	39 (14.3%)	26.81	3.49	0.00
DeepSeek-VL-7B	53.54 0.00 53.53 20.44	0.00 0.00 19.30 0.00	33.85 35.35 42.69 90.40	0.00 21.59 33.01 0.00	49.78 0.00 5.80 16.83	27.69 0.00 0.00 42.60	50.71 40.14 0.00 9.44	0.00 7.80 51.36 9.78	6.00 0.00 0.00 11.97	9.41 0.00 50.71 65.05	180 (66.2%)	64 (23.5%)	13.13	5.75	0.00
InternVL2.5-8B	59.96 0.00 30.13 18.18	4.86 0.00 28.37 0.00	24.93 26.68 46.05 99.60	0.00 17.74 16.95 0.00	38.08 0.00 7.82 10.57	35.39 0.00 0.00 85.90	57.54 48.81 0.00 33.52	0.00 8.06 54.99 9.71	7.76 0.00 0.00 16.91	12.46 0.00 74.49 57.17	183 (67.3%)	71 (26.1%)	25.20	13.09	0.00
Vitron-V1	47.64 0.00 50.00 19.07	3.90 66.60 28.14 36.70	51.58 39.47 22.28 51.38	2.30 8.19 23.52 55.85	35.66 58.53 0.00 4.70	4.81 82.72 44.96 69.26	39.78 25.13 0.00 15.34	0.00 22.24 52.11 19.12	13.30 14.63 71.89 24.48	13.81 0.00 64.20 59.07	252 (92.6%)	62 (22.8%)	30.13	7.65	4.59
MoE-LLAVA-Phi2-2.7B-4e-384	50.47 0.00 51.13 15.69	1.90 0.00 20.92 0.00	32.31 33.98 26.99 50.40	0.00 19.87 35.40 0.00	42.52 0.00 80.80 13.71	11.84 0.00 0.00 84.05	50.88 41.79 0.00 8.70	0.00 8.62 52.00 12.57	3.80 0.00 0.00 15.95	22.11 0.00 52.73 51.52	180 (66.2%)	55 (20.2%)	12.55	5.47	0.00
mPLUG-Owl2-LLaMA2-7b	52.53 0.00 51.60 15.27	0.00 0.00 23.60 0.00	26.00 29.01 41.66 60.20	0.00 18.67 27.08 0.00	36.72 0.00 86.80 9.00	12.35 0.00 0.00 80.10	44.03 31.75 0.00 8.88	0.00 9.39 51.67 12.14	0.60 0.00 0.00 17.48	20.88 0.00 42.51 70.10	177 (65.1%)	45 (16.5%)	12.21	4.60	0.00

on the diversity of its evaluation data, covering a wide range of fields and scenarios to assess different aspects of model capabilities. First, the dataset spans a variety of domains and disciplines, incorporating 28 major areas within both the physical sciences (e.g., Physics, Math, Geometry, Biology) and the social sciences (e.g., Humanities, Linguistics, History, Social). The evaluation of a generalist's skills and capabilities is categorized into universal modality-invariant

Table 3: Performance of part of multimodal generalists on image generation skills.

Model	Image Generation Skill (Avg within each #I-G Group)								Task Completion		Level Score on Image		
	#1 #9	#2 #10	#3 #11	#4 #12	#5 #13	#6 #14	#7 #15	#8	#Supported Task	#Winning-Specialist	Level-2	Level-3	Level-4
SoTA Specialist	18.70 53.16	45.40 16.47	33.77 25.33	16.30 43.93	4.86 20.35	24.00 67.44	99.29 36.11	15.06	/	/	/	/	/
SEED-LLaMA-14B	127.10 30.18	0.00 87.90	37.10 14.58	7.51 175.33	127.42 0.00	98.33 51.82	0.00 62.60	0.00	35 (77.8%)	0 (0.0%)	26.81	3.49	0.00
Emu2-32B	93.52 40.51	0.00 118.55	34.85 15.43	8.53 154.26	101.80 0.00	81.95 57.09	0.00 58.17	0.00	34 (75.6%)	2 (4.4%)	30.90	5.18	1.25
AnyGPT	158.21 28.88	0.00 108.06	40.47 14.91	10.30 193.39	117.21 0.00	115.91 53.02	0.00 64.21	0.00	36 (80.0%)	0 (0.0%)	23.10	1.29	0.00
LaVIT-V2 (7B)	79.79 46.40	0.00 89.78	31.35 15.79	11.87 161.54	149.78 0.00	59.23 50.18	0.00 51.68	0.00	36 (80.0%)	0 (0.0%)	29.50	3.71	0.00
NExT-GPT-V1.5	49.71 28.19	0.00 86.45	6.00 6.53	3.91 53.42	75.71 12.45	41.20 38.98	0.00 72.72	47.30	41 (91.1%)	0 (0.0%)	18.69	3.24	0.00
Vitron-V1	19.78 37.88	0.00 24.89	21.17 17.95	7.45 31.04	32.15 0.00	35.33 48.30	86.53 58.87	23.47	42 (93.3%)	3 (6.7%)	30.13	7.65	4.59

abilities and modality-specific skills. The modality-invariant abilities comprehensively include 12 categories, such as content recognition, commonsense knowledge, reasoning ability, causality discrimination, affective analysis, creativity, and innovation, etc. For modality-specific skills, we explicitly detail the main capabilities under both comprehension and generation for each modality, which correspond to the meta-tasks (skills) of our dataset.

5 Experiments

In this section, we conduct a comprehensive evaluation on `General-Bench`, from which we gain observations and jump to some conclusions.

5.1 Experimental Settings

For different models, we consistently follow the settings provided in their respective GitHub repositories, including model parameters and hyperparameters. We do not perform additional pre-training or fine-tuning. Each task and dataset comes with a predefined instruction prompt text. During evaluation, we use the same default prompt across all MLLMs to ensure fairness. The inference time varies across models. Smaller models complete evaluations within a few minutes, while larger models require significantly more time. On pure text-based NLP tasks, model inference is highly efficient; however, on video tasks, models demand more memory and have slower inference speeds. Our open-source codebase supports multi-GPU distributed inference, effectively accelerating the evaluation process. Also, we organize personnel into multiple groups to run models in parallel, further optimizing efficiency. For each task, we provide predefined evaluation scripts. Once the model generates outputs, the scripts are used to evaluate performance systematically.

5.2 Main Evaluation Results

We note that all the generalists run the evaluation on our `General-Bench` data set under a zero-shot setting. The overall results of part of the models on image comprehension and generation are presented in Table 2 and Table 3, respectively; Due to space limitations, we move the rest main results in Appendix §C.2. video results are shown in Table 11; audio results are shown in Table 12; 3D results are shown in Table 13; The results of all generalists on NLP tasks are shown in Table 14. The complete performing scores of all MLLMs across all tasks and datasets are presented in Appendix §C. Overall, we have the following observations.

Observation-1: Lack of task support. From these results, the first observation is that the vast majority of MLLMs exhibit a lack of support for a wide range of tasks in our benchmarks. Even models like OpenAI's GPT-4V and GPT-4o, which achieve top rankings on many existing MLLM benchmarks and leaderboards (Li et al., 2023b; Liu et al., 2024a), fail to demonstrate satisfactory task support on our benchmark. Specifically, GPT-4V and GPT-4o support only 177 out of 271 image comprehension tasks (65.1%). Among open-source models, InternVL2.5-8B achieves a task support rate of 71% for image comprehension tasks, outperforming GPT-4V and GPT-4o. For other modalities—such as video, audio, and 3D—the task-supporting rates are much less. Only Vitron-V1 supports over 90% of image tasks, and Sa2VA-8B achieves 72.2% supporting rate in the video comprehension group. This highlights a pervasive issue: current MLLMs require significant improvements in their architectural design to support as many tasks as possible.

Observation-2: Few generalists surpass the SoTA specialist. Also, we can notice that there are few models capable of surpassing the SoTA generalist. Overall, the tasks and skills that various MLLMs can surpass the SoTA specialists

are quite few. As seen, closed-sourced models (e.g., GPT-4V, GPT-4o, Gemini-1.5, and Claude-3.5) have the highest winning rate, with over 30% The best open-sourced Qwen2-VL-72B achieves a rate of 36.4% image comprehension by surpassing SoTA specialists. In other modalities such as video, audio, 3D, and language, the chances to surpass SoTA specialists are much lower. If an MLLM cannot outperform the SoTA specialist, it implies that the foundational conditions of cross-task/ability synergy for these MLLMs to become multimodal generalists are not met.

Observation-3: Focus more on content comprehension than supporting generation. For instance, GPT-4V and GPT-4o achieve better results than the SoTA specialist in certain skills within image comprehension tasks, and this improvement is significantly more pronounced than that of other models. However, GPT-4V and GPT-4o are limited to image comprehension tasks and provide zero support for image generation tasks. It is thus evident that GPT-4V and GPT-4o are not well-rounded multimodal generalists.[2] This trend becomes even more evident in other modalities. A significantly higher number of MLLMs support multimodal understanding compared to those supporting generation. Furthermore, the rate at which MLLMs surpass SoTA specialists in multimodal understanding benchmarks is much higher than in multimodal generation benchmarks. We emphasize that this imbalance reflects a critical limitation in the capability building of current multimodal generalists.

Observation-4: Insufficient support for all modalities. We also found that many MLLMs are unable to support all modalities simultaneously. Moreover, the vast majority of existing MLLMs are predominantly focused on understanding or generating image-based modalities. In contrast, much less attention has been devoted to video, audio, and 3D modalities (attention: image $>$ video $>$ 3D $>$ audio), with relatively few multimodal generalists addressing these areas. Most MLLMs, including the strongest ones, primarily handle image and language tasks, offering little to no support for other modalities. The completeness of support across various modalities and functionalities is insufficient for existing MLLMs to qualify as true multimodal generalists. We emphasize that to be considered a multimodal generalist, a model must be capable of understanding and generating signals from as many modalities as possible simultaneously.

Observation-5: Multimodality does NOT really enhance language. The ideal multimodal generalists should enable mutual enhancement across modalities. Unfortunately, our experimental results (as shown in Table 14) reveal that none of the current MLLMs provide any improvements in NLP tasks. Although various MLLMs achieve certain scores on NLP tasks, none of them surpass the performance of SoTA specialists in NLP. Furthermore, the performance gap between MLLMs and SoTA specialists in NLP tasks is larger than the gap observed in other modalities. While certain relevant research suggests that models, such as Vicuna, Qwen2, and LLaMA, trained with multimodal data (e.g., images) can also improve NLP tasks, such improvement has not yet enabled models to outperform SoTA NLP specialists on core language tasks. Our large-scale evaluation shows they still fall short of outperforming fine-tuned language specialists. We hypothesize that existing MLLMs, despite utilizing language-centered LLMs as their core, have significantly weakened their language capabilities due to an excessive focus on training and fine-tuning on non-language modalities. This trade-off not only undermines their language understanding but also fails to leverage multimodal information to enhance language-related tasks.

[2] It would thus be more rational to claim the current OpenAI GPT-4V/4o series as partial generalists, or visual generalists.

5.3 More Analyses and Discussions

Due to space limitations of the main article, we show much more experimental results as well as the analyses and discussions in Appendix, including:

- The main results of the multimodal generalists on other modalities, in §C.2;
- The complete results of models on specific tasks, from §C.6 to §C.10;
- The General-Level scores of leaderboards, in §C.3;
- Capability of tasks and modalities breakDown, in §C.4;
- Analysis and discussion on synergy, in §C.5;
- Future work of this project, in §D.

6 Conclusion

Inspired by the concept of capability levels defined in the autonomous driving industry, we propose `General-Level`, a framework that evaluates and categorizes the capabilities of existing MLLMs through a 5-level hierarchical rating mechanism based on their ability to maintain synergy across comprehension, generation, and multimodal interactions. `General-Level` provides a structured methodology to assess MLLMs across diverse tasks, modalities, and synergies, particularly in comprehension and generation. To support this evaluation, we further present `General-Bench`, a large-scale multimodal benchmark that spans a wide spectrum of tasks (702), modalities (language, image, video, audio, 3D, etc.), domains (29), and original formats with 325,800 instances. By benchmarking over 100 popular LLMs/MLLMs, we uncover the current capability limitations and provide a clear ranking of generalist performance. We hope the `General-Level` and `General-Bench` in this study will propel the community to develop next-generation multimodal foundation models to achieve more sophisticated, general-purpose multimodal intelligence.

Impact Statement

This work adheres to a rigorous ethical framework to ensure the responsible development, evaluation, and deployment of multimodal generalists. Below, we elaborate on the key ethical considerations. These measures ensure that General-Bench serves as a responsible and inclusive benchmark, contributing to the sustainable and equitable development of multimodal AI systems.

Privacy and Data Protection. The benchmark and evaluation process ensure strict compliance with privacy regulations. All tasks and datasets used in General-Bench are carefully curated to exclude personally identifiable information (PII). To safeguard privacy, any data derived from public sources is anonymized, and sensitive content is filtered out. Our procedures align with relevant data protection standards, such as GDPR and CCPA, emphasizing our commitment to ethical research practices.

Data Collection. The dataset for General-Bench is built using publicly available resources or through collaborations with contributors who explicitly consented to their data being included. Data collection protocols are designed to prioritize ethical sourcing, ensuring that contributors understand their rights, including the ability to withdraw their data at any time. This ensures transparency and fairness throughout the dataset construction process.

Annotator Compensation. Human annotators play a crucial role in ensuring the high quality of the General-Bench dataset. We engage well-trained annotators, including postgraduate students and crowdsourcing professionals, and provide fair compensation for their work. Annotators are either volunteered to contribute, or paid based on the estimated time required to complete specific tasks. All are signed to give their best efforts in data annotation and model implementation to ensure the work quality.

Bias and Fairness. Recognizing the potential biases in AI systems, we take active measures to analyze and mitigate biases related to gender, ethnicity, language, and other sociocultural factors within the dataset and evaluation tasks. Diverse and representative data collection practices are employed across multiple modalities and languages. While we acknowledge that complete eradication of bias is challenging, we strive to identify and address biases as the benchmark evolves.

Intellectual Property Protection. All datasets and tasks included in General-Bench respect intellectual property rights. Data collected from external sources is fully repurposed and modified, and is used under proper licensing agreements, ensuring compliance with intellectual property laws. Open-sourced models are strictly used according to their licenses. Models evaluated via APIs are handled according to their respective terms of use, and no proprietary content is redistributed without permission.

Misuse Potential. We are aware of the potential risks associated with misuse of multimodal intelligence technologies, such as applications in surveillance or the manipulation of public opinion. To mitigate such risks, we have developed guidelines to encourage ethical use. These guidelines emphasize the importance of transparency, accountability, and consent in any application or further development of the technologies evaluated in this work.

Accessibility and Inclusivity. In alignment with our commitment to fostering inclusivity in the AI research community, all code, tasks, and datasets related to General-Bench are openly available. This ensures that researchers from diverse backgrounds and varying resource levels can equally contribute to, and benefit from, advancements in multimodal generalist research.

Evaluations and Performance. We clarify that the performances of all models reported in this paper—including both specialists and generalists—are influenced by the specific testing environment. This includes factors such as the size and content of the dataset, as well as the parameters used in the reproduced code. As we continuously update the dataset, the evaluation results presented in this paper may differ from those obtained in future versions. We emphasize that such differences are considered reasonable and expected deviations, and should not raise any concerns. Our leaderboard is open and under active development, and we warmly welcome participation from external practitioners.

References

OpenAI. Introducing chatgpt. 2022a.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the NeurIPS*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training

with frozen image encoders and large language models. In *Proceedings of the ICML*, pages 19730–19742, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023a.

OpenAI. Gpt-4 technical report. 2022b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023a.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *CoRR*, abs/2305.11000, 2023a.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning*, 2024a.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438, 2024a.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation. *arXiv preprint arXiv:2311.16511*, 2023a.

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024a.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. 2024a.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. *arXiv preprint arXiv:2312.10032*, 2023a.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024a.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023a.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024a.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024a.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024a.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024b.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023a.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2024a.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*, 2024b.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023b.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024c.

Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8, 2024b.

Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024c.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the ICML*, pages 12888–12900, 2022.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2023.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024a.

Anthropic Team. The claude 3 model family: Opus, sonnet, haiku. *preprint*, 2024. URL https://api.semanticscholar.org/CorpusID:268232499.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024b.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024d.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024c.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, pages 216–233, 2024a.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023b.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580, 2023.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023c.

Kaihang Pan, Siliang Tang, Juncheng Li, Zhaoyu Fan, Wei Chow, Shuicheng Yan, Tat-Seng Chua, Yueting Zhuang, and Hanwang Zhang. Auto-encoding morph-tokens for multimodal llm. *arXiv preprint arXiv:2405.01926*, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521, 2022.

Jingkun Ma, Runzhe Zhan, Derek F Wong, Yang Li, Di Sun, Hou Pong Chan, and Lidia S Chao. Visaidmath: Benchmarking visual-aided mathematical reasoning. *arXiv preprint arXiv:2410.22995*, 2024.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024e.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models. *CoRR*, abs/2406.16020, 2024b.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger B. Dannenberg, Wenhu Chen, Gus Xia, Wei Xue, Si Liu, Shi Wang, Ruibo Liu, Yike Guo, and Jie Fu. MARBLE: music audio representation benchmark for universal evaluation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. *CoRR*, abs/2410.19168, 2024.

Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T^3bench: Benchmarking current progress in text-to-3d generation. *CoRR*, abs/2310.02977, 2023.

Junjie Zhang, Tianci Hu, Xiaoshui Huang, Yongshun Gong, and Dan Zeng. 3dbench: A scalable 3d benchmark and instruction-tuning dataset. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 1706–1714, 2024b.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *CoRR*, abs/2311.16103, 2023.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis. *CoRR*, abs/2405.21075, 2024b.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: A comprehensive benchmark for multi-task long video understanding. *CoRR*, abs/2406.04264, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206, 2024f.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXVIII*, pages 179–195, 2024d.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21807–21818, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *CoRR*, abs/2306.09265, 2023b.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13084–13094, 2024.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, and Huaxiu Yao. MMIE: massive multimodal interleaved comprehension benchmark for large vision-language models. *CoRR*, abs/2410.10139, 2024b.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms. *CoRR*, abs/2407.01509, 2024.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? *CoRR*, abs/2408.13257, 2024c.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Song Dingjie, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jianquan Li, Xiang Wan, and Benyou Wang. Mllm-bench: Evaluating multimodal llms with per-sample criteria, 2024.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024b.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. *CoRR*, abs/2406.09411, 2024c.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *CoRR*, abs/2404.18532, 2024.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *CoRR*, abs/2311.17092, 2023c.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *CoRR*, abs/2406.16860, 2024a.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: multimodal multi-image understanding for evaluating large vision-language models. *CoRR*, abs/2408.02718, 2024b.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024b.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, Dongfu Jiang, Xuan He, Yuan Liu, Hexiang Hu, Xiang Yue, and Wenhu Chen. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *CoRR*, abs/2410.10563, 2024e.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024b.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024a.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024g.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024b.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, et al. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*, 2023.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023d.

Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023c.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023d.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o:

One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2025.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024a.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. arxiv 2024. *arXiv preprint arXiv:2309.04669*.

Boyang Zheng, Jinjin Gu, Shijun Li, and Chao Dong. Lm4lv: A frozen large language model for low-level vision tasks. *arXiv preprint arXiv:2405.15734*, 2024.

Yikang Zhou, Tao Zhang, Shilin Xu, Shihao Chen, Qianyu Zhou, Yunhai Tong, Shunping Ji, Jiangning Zhang, Xiangtai Li, and Lu Qi. Are they the same? exploring visual correspondence shortcomings of multimodal llms, 2025.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024d.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024b.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv*, 2025.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36: 20482–20494, 2023.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2025.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023b.

Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1357–1366, 2024b.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024b.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. *arXiv preprint arXiv:2306.10900*, 2023e.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313, 2024.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108, 2023.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*, 2024b.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. Audiogpt: Understanding and generating speech, music, sound, and talking head. *CoRR*, abs/2304.12995, 2023.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355, 2023.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Xinyu Wang, Bohan Zhuang, and Qi Wu. Modaverse: Efficiently transforming modalities with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26606–26616, 2024e.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *CoRR*, abs/2205.15868, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

On Path to Multimodal Generalist: General-Level and General-Bench

Full-version Paper: https://arxiv.org/abs/2505.04620

Project Page: https://generalist.top

Leaderboard: https://generalist.top/leaderboard

Benchmark: https://huggingface.co/General-Level

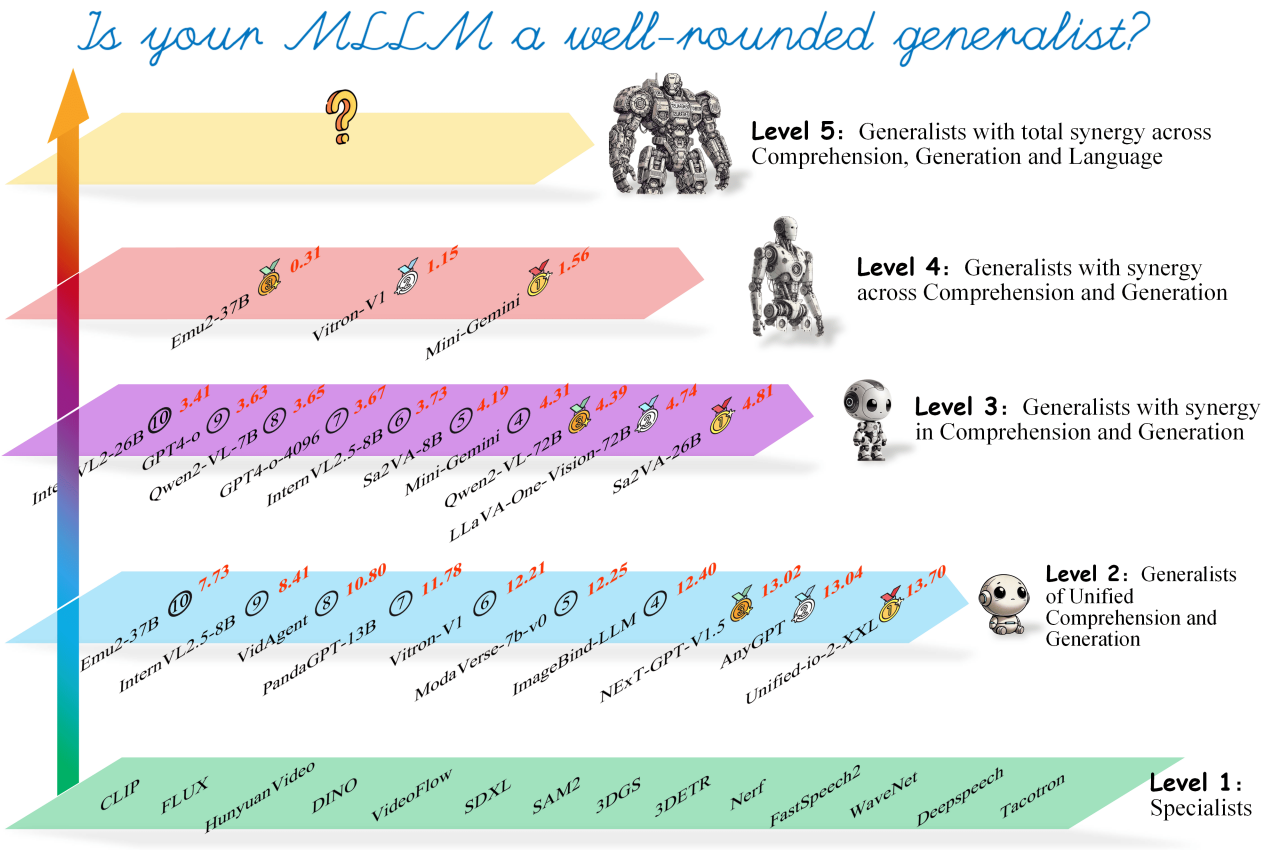


Figure 4: Leaderboard of multimodal generalists over **`General-Level`** (only top-performing ones shown here).

Table of Contents

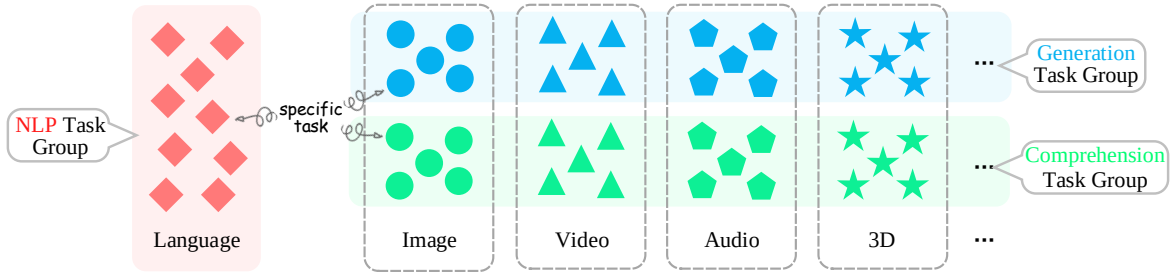


Figure 5: We categorize tasks of various modalities into Comprehension group, Generation group and NLP group. Each colored stylish symbol represents a specific task of a certain modality.

A Extension on General-Level Scoring

A.1 Scoring Specification

Assume we have a benchmark of various modalities and tasks, where we can categorize tasks under these modalities into the Comprehension group and the Generation group, as well as the language (i.e., NLP) group, as illustrated in Figure 5. Then, we define the scoring specification of `General-Level` as in Table 1.

Also, except for Level-1 and Level-5, when calculating S_2, S_3, and S_4, we consider a reasonable approach when handling different modalities. First, we calculate the specific score component S_k^i of a generalist in the i-th modality (assuming there are N modalities in total) for the score S_k. This modality-specific component can accurately reflect the model's Level-k capability in the i-th modality. Next, by decomposing each score into its components across different modalities, we sum the components of each modality with equal weights to obtain the overall score for each level.

$$S_k = \sum_{i}^{N} \frac{1}{N} S_k^i$$

The advantage of this method is that it reduces the bias introduced by the number of tasks in different modalities. For example, in our benchmark, image-related tasks (especially comprehension-type tasks) are overwhelmingly more numerous compared to other modalities, such as audio tasks. Therefore, two generalists with similar capability levels, say one for image tasks and the other for audio tasks, would have a higher S_k score for the image-generalist over the audio-generalist, due to the larger number of image tasks. This discrepancy is unrealistic and contrary to our core idea for evaluating multimodal generalists. To eliminate the bias caused by the number of tasks within each modality, we propose the above calculation method, which treats the capabilities of different modalities equally. Meanwhile, this method also prioritizes generalists that can support more modalities. For instance, a model that supports more modalities will certainly have a higher overall score compared to a generalist that supports only one modality.

This scoring method ensures that as an MLLM climbs to higher levels, its scores progressively decrease, which should indicate the increasing difficulty of advancing levels. Climbing from level n to level $n+1$ requires specific capabilities, i.e., demonstrating sufficient synergy capability associated with that level, which we highlight as critical factors in Table 1. Within the same level, to achieve a higher score, a model must: 1) support as many tasks and modalities as possible, and simultaneously 2) achieve the highest possible performance on individual tasks.

A.2 Scoring Relaxation

A central aspect of our `General-Level` framework lies in how synergy effects are computed. According to the standard understanding of the 'synergy' concept, e.g., *the performance of a generalist model on joint modeling of tasks A and B (e.g., $P_\theta(y|A,B)$) should exceed its performance when modeling task A alone (e.g., $P_\theta(y|A)$) or task B alone (e.g., $P_\theta(y|B)$).* However, adopting this approach poses a significant challenge that hinders the measurement of synergy: there is no feasible way to establish two independent distributions, $P_\theta(y|A)$ and $P_\theta(y|B)$, and a joint distribution $P_\theta(y|A,B)$. This limitation arises because a given generalist model has already undergone extensive pre-training and fine-tuning, where tasks A and B have likely been jointly modeled. It is impractical to retrain such a generalist to isolate the learning and modeling of tasks A or B independently in order to derive these distributions. Otherwise, such an approach would result in excessive redundant computation and inference on the benchmark data.

To simplify and relax the evaluation of synergy, we introduce a key assumption in the scoring algorithm:

> *Theoretically, we posit that the stronger a model's synergy capability, the more likely it is to surpass the task performance of SoTA specialists when synergy is effectively employed. Then, we can simplify the synergy measurement as: if a generalist outperforms a SoTA specialist in a specific task, we consider it as evidence of a synergy effect, i.e., leveraging the knowledge learned from other tasks or modalities to enhance its performance in the targeted task.*

By making this assumption, we avoid the need for direct pairwise measurements between 'task-task', 'comprehension-generation', or 'modality-modality', which would otherwise require complex and computationally intensive algorithms.

A.3 Properties of General-Level

The `General-Level` framework possesses several important attributes that play a critical role in supporting the hierarchical classification and ranking of MLLMs. These properties are also well-grounded in mathematical theory.

Property-1: Independence from Peer Generalists In our scoring framework, the scores of any generalist depend solely on the dataset and the reference scores of SoTA specialists, without relying on the scores of other tested generalists. These two components are entirely independent. The dataset defines the specific tasks, while the specialists provide baseline reference scores used for the calculation of the experimental generalists' scores. This property ensures that the evaluation of generalists is free from interdependence, maintaining objectivity and fairness among all systems participating in the ranking.

Property-2: Monotonicity Across Levels Generally, if a generalist is rated at the highest level-k, it is expected to achieve scores at all levels from 2 to k. We further expect that as the level increases, the corresponding scores for the generalist will decrease, i.e., $S_{k-1} > S_k$. This is a reasonable and realistic requirement, as higher levels impose stricter demands on the generalist's capabilities, naturally leading to lower scores for the same model. Below, we provide proof that the scoring algorithm of `General-Level` framework mathematically guarantees the strictly monotonic score decline across levels.

► The proof for $S_3 \leq S_2$

$$
\begin{aligned}
S_3 =& \frac{1}{2}\left(S_G + S_C\right) \\
=& \frac{1}{2}\left(\frac{1}{M}\sum_{i=1}^{M}\begin{cases}\sigma_i^C & \text{if } \sigma_i^C \geq \sigma_{sota}^C \\ 0 & \text{otherwise}\end{cases} + \frac{1}{N}\sum_{j=1}^{N}\begin{cases}\sigma_j^G & \text{if } \sigma_j^G \geq \sigma_{sota}^G \\ 0 & \text{otherwise}\end{cases}\right) \\
\leq& \frac{1}{2}\left(\frac{1}{M}\sum_{i=1}^{M}\sigma_i^C + \frac{1}{N}\sum_{j=1}^{N}\sigma_i^G\right) \\
=& S_2
\end{aligned}
$$

► The proof for $S_4 \leq S_3$

Suppose:

$$
S_G = \frac{1}{M}\sum_{i=1}^{M}\begin{cases}\sigma_i & \text{if } \sigma_i \geq \sigma_{sota} \\ 0 & \text{otherwise}\end{cases}
$$

$$
S_C = \frac{1}{N}\sum_{j=1}^{N}\begin{cases}\sigma_j & \text{if } \sigma_j \geq \sigma_{sota} \\ 0 & \text{otherwise}\end{cases}
$$

According to *Cauchy-Schwarz Inequality*, let's represent

$$
\left(\frac{S_C + S_G}{2}\right)^2 \geq \left(\frac{2S_C S_G}{S_C + S_G}\right)
$$

Expanding this,

$$
\frac{(S_C + S_G)^2}{4} \geq \frac{2S_C S_G}{S_C + S_G}
$$

Multiplying both sides by $4(S_C + S_G)$,

$$(S_C + S_G)^3 \geq 8S_C S_G(S_C + S_G)$$

Simplifying further

$$S_C^3 + S_G^3 \geq 2S_C S_G(S_C + S_G)$$

This factorizes to

$$(S_C - S_G)^2(S_C + S_G) \geq 0\,.$$

Finally, we have

$$\begin{aligned} S_4 &= \frac{2S_C S_G}{S_C + S_G} \\ &\leq \frac{1}{2}(S_C + S_G) \\ &= S_3\,. \end{aligned}$$

► The proof for $S_5 \leq S_4$

We have

$$w_L = \frac{S_L}{S_{\text{total}}}\text{, where}$$
$$S_L = \frac{1}{T}\sum_{k=1}^{T}\begin{cases}\sigma_k & \text{if } \sigma_k \geq \sigma_{\text{sota}} \\ 0 & \text{otherwise}\end{cases}$$

which means,

$$w_L \leq 1\,.$$

Then

$$\begin{aligned} S_5 - S_4 &= S_4 * w_L - S_4 \\ &= S_4 * (w_L - 1) \\ &\leq 0 \end{aligned}$$

Thus,

$$S_5 - S_4 \leq 0$$

Property-3: Encouraging Rich and Balanced Multimodal Task Support.

► **More Task, The Better.** A good multimodal evaluation system should not only reward models for achieving higher scores on individual tasks and surpassing SoTA specialists but also incentivize a trend where multimodal generalists support as many diverse multimodal tasks as possible. This is a reasonable expectation, as an ideal multimodal generalist should inherently support a broader range of modalities and tasks. The scoring algorithm of our `General-Level` framework aligns with this objective. For instance, in the case of level-2 scoring:

$$S_2 = \frac{1}{M+N}\sum_{i=1}^{M+N}\sigma_i\,,$$

a model that achieves nonzero scores across a greater number of modalities and tasks will naturally obtain a higher average score, thereby ranking higher within the same level.

► **More Balance, The Better.** Moreover, our scoring algorithm also promotes models that achieve more balanced performance across tasks. For example, in the case of level-4 scoring, consider the following scenarios:

1) Model A achieves SoTA specialist performance on X tasks in the comprehension category but only Y tasks (where $X \gg Y$) in the generation category.
2) Model B achieves SoTA specialist performance on X tasks in both the comprehension and generation categories.

According to the properties of the harmonic mean inequality, $S_4^A < S_4^B$.

► **The proof for $S_4^A < S_4^B$ when $X \gg Y$ in level-4**

Extreme Assumptions:
- For Model A, the X tasks in the comprehension group have scores of $\sigma_C^A = 1$, and the Y tasks in the generation group have scores of $\sigma_G^A = 1$, while all other scores are 0.
- For Model B, both comprehension and generation groups have X tasks with scores of $\sigma_C^B = 1$ and $\sigma_G^B = 1$, while all other scores are 0.

Model-A Scores:
For Model A, the comprehension and generation scores are:

$$S_C^A = \frac{X}{M}, \quad S_G^A = \frac{Y}{N}.$$

The overall score for Model A is:

$$S_4^A = \frac{2 \cdot S_C^A \cdot S_G^A}{S_C^A + S_G^A} = \frac{2 \cdot \frac{X}{M} \cdot \frac{Y}{N}}{\frac{X}{M} + \frac{Y}{N}} = \frac{2XY}{XN + YM}.$$

Model-B Scores:
For Model B, both comprehension and generation groups have X tasks with scores of 1, so:

$$S_C^B = \frac{X}{M}, \quad S_G^B = \frac{X}{N}.$$

The overall score for Model B is:

$$S_4^B = \frac{2 \cdot S_C^B \cdot S_G^B}{S_C^B + S_G^B} = \frac{2 \cdot \frac{X}{M} \cdot \frac{X}{N}}{\frac{X}{M} + \frac{X}{N}} = \frac{X^2}{XN + XM}.$$

Comparison:
We need to compare:

$$\frac{2XY}{XN + YM} \quad \text{and} \quad \frac{X^2}{XN + XM}.$$

Given $X \gg Y$, it follows that:

$$\frac{2XY}{XN + YM} < \frac{X^2}{XN + XM}.$$

Thus, $S_4^A < S_4^B$.

Through the above mathematical analysis, we have proven that under the same task distribution, the uneven generation score distribution of Model A results in its level-4 score being lower than that of Model B. This ensures that models with more balanced performance across comprehension and generation are ranked higher.

Property-4: Dynamic Update on Benchmarking and Specialists Finally, we observe an important point: the more tasks included in the benchmark used to evaluate models, the more accurate and objective the resulting evaluations and conclusions. This requirement for the evaluation benchmark to have dynamic properties aligns well with real-world needs. In practice, new tasks, data, and even new modalities are constantly being introduced, and a generalist should be capable of covering these newly added tasks and functionalities. Accordingly, in our evaluation system, we allow the benchmark to evolve dynamically, such as by adding new tasks under various modalities and categories. Once new tasks are added, we update the scores and rankings of all tested generalists to reflect the expanded benchmark.

On the other hand, we also allow updates to the SoTA specialist models timely for each task, as scoring at higher levels is anchored to the performance of the SoTA models. This is a reasonable act, as specialists are continually being developed and improved. Once a baseline specialist advances, generalists must also improve to remain competitive, or risk being surpassed. Thus, in `General-Level` framework, the scores corresponding to SoTA specialists are subject to periodic updates. Also, we dynamically and regularly update the scoring and ranking of all generalists to ensure the evaluation remains accurate and reflective of the current state of the field.

A.4 Receipt to Leveling Upper in General-Level

Here we provide a guideline to help better understand how to achieve higher levels in `General-Level` framework.

Level-1→Level-2: Supporting as many tasks and functionalities as possible. Transitioning from specialists to generalists requires making the system compatible with various task modeling paradigms, i.e., supporting diverse modality types and input formats, as well as handling a wide range of model types and output formats (whether for comprehension and/or generation). Currently, the most popular and widely adopted practice is to use an LLM as the backbone/intelligence medium, integrating various specialists to build generalists. There are two primary implementation strategies.

First, agent-based generalists (Wu et al., 2023b; Shen et al., 2023). In this approach, the LLM acts as a task scheduler and dispatcher, facilitating message passing through hard integration (explicit text). This is essentially a pipeline architecture. However, since gradient propagation across the entire system is not feasible, this method is prone to error propagation. The performance upper bound of generalists built with this approach is equivalent to the SoTA specialists for all supported tasks, primarily due to the lack of features, information sharing, and limited task collaboration.

Second, end-to-end generalists (Liu et al., 2023c; Li et al., 2023a; Zhu et al., 2023a). In this type, the entire system is constructed as a continuous joint model, allowing for full-stack updates via gradient propagation. The most common architecture in this category uses an LLM as the backbone, achieving soft integration of various encoders and decoders through input tokenization and feature embedding, combined with overall fine-tuning.

Level-2 → Level-3: Generalists achieving as stronger synergy and cross as many tasks as possible. To advance from a vanilla generalist to Level-3, the system must demonstrate cross-task synergy capabilities, enabling at least two tasks (regardless of whether both involve comprehension, generation, or one involves comprehension while the other involves generation) to share features and achieve mutual performance improvements. The most direct method to realize cross-task synergy is through multi-task joint training. Specifically, during joint learning, the system must ensure it can maintain task-shared/persistent common features while preserving each task's specific features without degradation, e.g., Vitron (Fei et al., 2024a). Moreover, the model must support synergy across as many tasks as possible and ensure that the synergy effect is significant enough to achieve higher evaluations at Level-3.

Level-3→Level-4: Generalists in unified comprehension and generation capability with synergy in between. To advance to Level-4, generalists must first achieve unified comprehension and generation capabilities, regardless of whether they support a single modality (non-NLP) or multiple modalities. At the same time, the system must meet the requirement that its capabilities in comprehension and generation synergize and enhance one another. Generally speaking, compared to acquiring comprehension capabilities, obtaining generation capabilities at the technical level is relatively more challenging. For instance, the visual comprehension abilities of most visual LLMs tend to be significantly stronger than their visual generation capabilities. If a generalist can score at Level-4, it indicates that the system not only possesses strong comprehension capabilities but also maintains these capabilities while further learning and training its generation abilities. To achieve this, Morph-Token (Pan et al., 2024) introduces a disentangling visual reconstruction loss for generation learning to avoid interference with the comprehension learning loss.

Level-4→Level-5: Generalists achieving cross-modal synergy with abductive reasoning ability. Achieving Level-5 represents the ultimate goal for generalists, where features, knowledge, and even intelligence learned from tasks in certain modalities can (to varying degrees) transfer to tasks in other supported modalities. Currently, most multimodal generalists are limited by architectural developments, primarily enabling language intelligence to support intelligence in other modalities (as illustrated in Figure 1). However, to truly achieve Level-5, synergy must exist across all modalities. For instance, in the current MLLM community, this would require MLLMs to enhance performance on NLP tasks as well, while most of the MLLMs perform unsatisfactorily in NLP tasks. From a technical perspective, generalists must be capable of abductive reasoning, i.e., the ability to infer and generalize across everything. Also, they need to ensure modality-agnostic context consistency during reasoning.

B Extension on General-Bench Dataset

This part provides an extension to our `General-Bench` dataset.

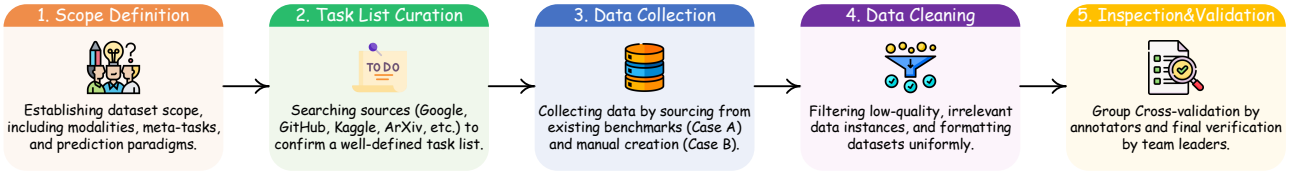


Figure 6: An illustration of the data construction pipeline of `General-Bench`.

B.1 Data Construction

B.1.1 Design Criterion

As previously noted, the current benchmarks that rank MLLMs based solely on their performance have significant limitations, which hinder the encouragement of MLLMs to evolve toward becoming more capable multimodal generalists. Primarily, nearly all existing benchmarks focus on evaluating MLLMs' capabilities in visual modalities, particularly images, while significantly neglecting tasks in other modalities such as video, audio, 3D, etc. Moreover, they often assume that MLLMs already possess satisfied NLP capabilities, thus omitting evaluations in language.

Secondly, these benchmarks tend to simply convert free-form predictions into fixed QA format of pre-defined choices—essentially a compromise that reflects the current limitations of MLLM capabilities—allowing many tasks that MLLMs cannot produce in specific formats to still be executed. We believe that a genuine multimodal generalist should support tasks in their original formats. Furthermore, most benchmarks only assess MLLMs' understanding of visual information; however, a multimodal generalist should inherently possess a wide range of capabilities beyond mere comprehension, such as generation, editing, etc. Therefore, we expect to construct a benchmark that possesses these characteristics:

- Covering as broad a range of tasks, skills and modalities as possible.
- Encompassing both comprehension and generation of tasks.
- Including a rich diversity of tasks across various scenarios and domains.
- Preserving the original task-prediction formats.
- Timely maintaining and expanding the dataset dynamically.

B.1.2 Construction Process

The construction of our `General-Bench` dataset follows a structured 5-step process to ensure both comprehensiveness and quality. Figure 6 presents the data construction pipeline.

Step-1: Defining Scope and Range. We begin by conducting a series of panel discussions to establish the scope of the dataset. This involves determining the modalities to include, identifying the core general skills (meta-tasks), and specifying the prediction paradigms to address. These discussions help outline a comprehensive framework for the dataset, ensuring that it accommodates diverse tasks and capabilities required for evaluating multimodal generalists.

Step-2: Curating Task List. Based on the defined scope, we curate a comprehensive task list by systematically searching various sources, including Google, GitHub, Kaggle, ArXiv, and PaperWithCode, etc. For each task, we specify its input-output targets, select appropriate evaluation metrics, and also identify SoTA specialists as reference points. This step ensures that each task is well-defined and aligned with existing SoTA practices.

Step-3: Collecting Data. Next, we start collecting the data instances. The data collection process is divided into two cases for handling two different scenarios:

- **Case A:** If the data could be sourced from existing benchmark datasets (only from their test sets), modifications are made to enhance diversity. We will show all the data sources of our benchmark in the following subsections. For textual data, rephrasing is done using ChatGPT. For non-textual modalities such as images, videos, and audio, semantically equivalent replacements are identified through retrieval or direct recording from relevant databases or websites.
- **Case B:** For tasks without available datasets or insufficient enough numbers of samples, we manually create instances. This involves crafting input-output pairs according to the task definition, running existing models to generate predictions, and performing manual verification and correction of the results.

We ensure that each task includes (at least) 500 data samples. Also, we ensure that all tasks faithfully retain their original

input-output prediction structure or format, i.e., not reformatted into QA-based multiple-choice questions.

Step-4: Data Filtering and Cleaning. After collecting datasets for all modalities and tasks, we proceed with data filtering and cleaning. First, we filter out low-quality instances, including those that do not align well with the task's evaluation purpose, lack target modality information, or fail to meet the defined prediction paradigms. For tasks where the number of instances is insufficient, we restart the data annotation process to supplement the required quantity. Afterward, we organize all data into a unified storage format according to the designed specifications. For example, textual data is standardized into JSON files with consistent naming conventions applied to all files.

Step-5: Data Inspection and Validation. Finally, we conduct a rigorous inspection and validation process to guarantee data quality and consistency. Annotators work in groups of three, independently reviewing the same instance. An instance is accepted only if all three annotators reach consensus. Finally, team leaders or supervisors conduct an additional round of verification to ensure the dataset meets the highest standards of consistency and accuracy.

B.2 Evaluation and Splitting

As each task follows the original format, our evaluation metrics vary in rich task types. For instance, we evaluate X-to-text generation tasks using BLEU/ROUGE/CIDEr scores, image segmentation tasks with mIoU for generating masks, and image generation tasks using FID, etc. Also, we design some mapping functions to standardize performance scores. In Appendix §B.4 we present the evaluation metrics as well as the mapping tricks in detail.

For most of the tasks, we maintain around 500 testing instances each. Considering that not all practitioners in the community may be interested in participating in the leaderboard—for example, some may simply wish to use our dataset for their research or publications—we propose dividing the test set for each task into a closed set and an open set. The closed set is reserved for leaderboard evaluations: only the input data is released, and users are required to submit their model's predicted outputs for centralized assessment. In contrast, the open set provides full access to both inputs and corresponding outputs, enabling practitioners to explore and utilize the data more freely. Each task's test set is split into closed and open subsets with a ratio of 2:3.

B.3 Leaderboard Re-Scoping

Given the large scale of our dataset, it would be highly costly for practitioners to run the entire dataset under our proposed General-Level evaluation protocol. Moreover, it's realized that most existing multimodal generalists (e.g., MLLMs) have not yet reached the level of capability required to cover a wide range of modalities and tasks, as envisioned in our framework. As a result, many current models may find it difficult to fully demonstrate their potential on our leaderboard. To improve usability and encourage broader participation, we further propose a graded structure for the leaderboard by dividing its scope into four levels of increasing difficulty:

- **Scope-A**: Full-spectrum leaderboard covering all modalities and tasks, designed for highly capable, general-purpose multimodal models. This scope has one leaderboard encompassing all levels in General-Level, making it the most challenging track. We further derive a full version and a quick version leaderboard for easier participation.
- **Scope-B**: Modality-specific leaderboards, each focusing on a single modality or partially joint modality, and designed for modality-wise generalists. This scope maintains 4 separate leaderboards, one per modality (except for language).
- **Scope-C**: Leaderboards focused on either comprehension or generation within a single modality. This scope includes 8 leaderboards: 2×4 for comprehension/generation across multimodal tasks, with a lower entry barrier for participation.
- **Scope-D**: Finer-grained, skill-level (task-cluster-specific) leaderboards within each modality, tailored for partial generalists. This scope includes a large number of specific leaderboards, offering the lowest difficulty for participation.

Figure 7 illustrates this design. Each leaderboard scope reflects a different level of difficulty, allowing practitioners to flexibly choose which leaderboard to participate in based on the capabilities of their models and the amount of resources they are willing to invest.

B.4 Evaluation Metrics

Metric List. Since all tasks in `General-Bench` retain their original task definitions without altering the output or prediction format, our evaluation methods vary according to the nature of different tasks and data. Table 4 summarizes the evaluation metrics and methods used across all tasks.

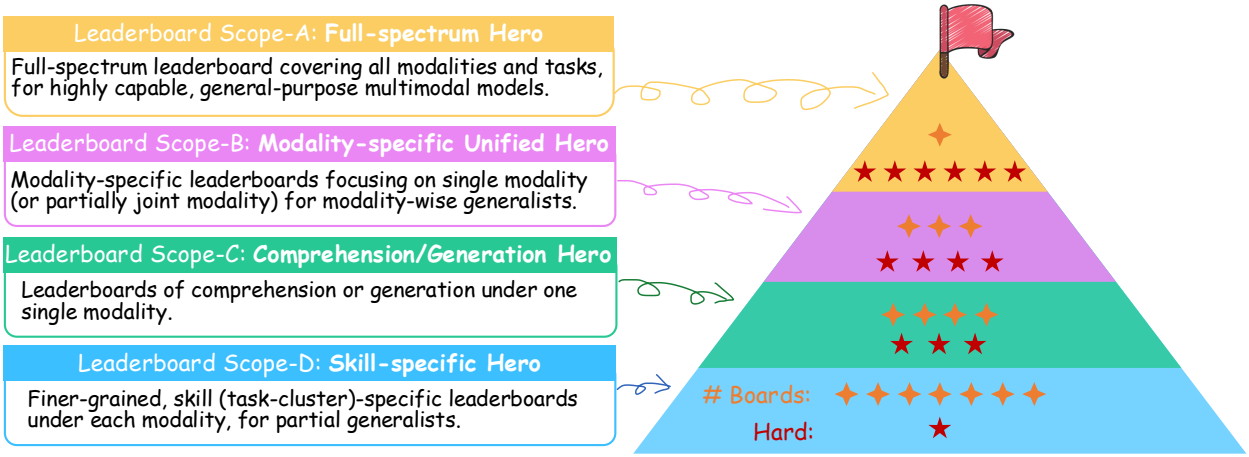


Figure 7: We reorganize **General-Bench** into 4 scopes, categorized by the level of participation difficulty for practitioners.

Table 4: Overview of all the evaluation metrics in `General-Bench`. ↑ means the higher the better performance, and vice versa for ↓.

#	Metric	Range	Calculation	Representative Tasks
• General				
1	Acc↑	[0,1]	Accuracy is defined as the ratio of correctly classified instances to the total number of instances.	Classification
2	Macro-Acc↑	[0,1]	Macro-Acc evaluates how well a model performs on average across all classes, regardless of class imbalance.	Event Relation Prediction
3	EM-Acc↑	[0,1]	Exact Match Accuracy evaluates the percentage of predictions that are exactly the same as their corresponding references.	QA, machine translation, or summarization
4	AP↑	[0,1]	AP, Average Precision, is a metric used to evaluate the performance of object detection tasks, reflecting the overall precision-recall trade-off across multiple thresholds.	Anomaly Detection
5	mAP ↑	[0,1]	mAP, Mean Average Precision, is the mean of Average Precision values across all queries or instances:	2D/3D Detection
6	F1↑	[0,1]	F1 score is the harmonic mean of Precision and Recall.	QA
7	Micro-F1↑	[0,1]	Micro-F1 score is the harmonic mean of the Micro-averaged precision and recall.	Classification
8	AUC↑	[0,1]	AUC is used in binary classification tasks and measures the area under the ROC curve. It represents the model's ability to distinguish between classes.	Image Generation
• Ranking-related				
9	R@k↑	[0,1]	R@k measures the Recall rate at the top k results in tasks like image retrieval, where the true positive must appear within the top k predicted results.	Image Scene Graph Parsing
10	AP@k↑	[0,1]	AP@k is the Average Precision calculated at an IoU threshold of k (k¡1). This metric is typically used when higher overlap between retrieved items and ground truth items is required.	Object Detection
11	mAP@k↑	[0,1]	mAP@k refers to the mean Average Precision where the Intersection over Union (IoU) threshold is set to k (k¡1).	Object Detection
12	EM@1↑	[0,1]	Exact Match at 1 evaluates the proportion of instances for which the model's top prediction exactly matches the correct answer.	3D Question Answering
13	ANLS↑	[0,1]	ANLS, Average Normalized Levenshtein Similarity, measures how well a model ranks items in a list based on their relevance to a query.	OCR
• Regression-related				
14	MAE ↓	[0,∞)	MAE, Mean Absolute Error, measures the average of the absolute differences between the predicted values and the actual values. It's typically used in regression tasks.	Object Counting
15	RMS ↓	[0,∞)	RMS, Root Mean Square, is a metric for regression tasks that measures the square root of the average squared differences between the predicted values and true values.	Image Depth Estimation
16	MSE ↓	[0,∞)	MSE, Mean Squared Error, is commonly used for regression tasks and measures the average squared differences between predicted values and actual values.	Object Matting
17	RMSE ↓	[0,∞)	RMSE, Root Mean Squared Error.	Time Series Prediction
• Text Generation-related				

#	Metric	Range	Calculation	Representative Tasks
18	BLEU-1↑	[0,1]	BLEU-1, Bilingual Evaluation Understudy (1-gram), calculates the precision of unigrams (individual words) in the generated text compared to the reference text(s).	Text Generation
19	BLEU-4↑	[0,1]	BLEU-4, Bilingual Evaluation Understudy (4-gram).	Text Generation
20	CodeBLEU↑	[0,1]	CodeBLEU is a metric designed to evaluate the quality of generated code by comparing it to reference code. CodeBLEU combines standard BLEU with additional features specific to code, such as syntax matching, data flow alignment, and weighted n-gram matching.	Code Generation
21	ROUGE-L↑	[0,1]	ROUGE, Recall-Oriented Understudy for Gisting Evaluation of Longest Common Subsequence, LCS, evaluates text generation tasks, which measures the overlap between the predicted text and the reference text.	Image/Video Captioning
22	ROUGE-1↑	[0,1]	ROUGE in 1-grams (single words).	Text Generation
23	CIDEr↑	[0,1]	Consensus-based Image Description Evaluation (CIDEr), evaluates the quality of generated sentences (e.g., image captions) by comparing them against a set of reference sentences.	Captioning
• Image-related				
24	PSNR↑	[0,∞)	PSNR, Peak Signal-to-Noise Ratio, measures the quality of a reconstructed or compressed signal, such as images or videos, compared to the original signal.	Image Desnowing
25	MS-SSIM↑	[-1,1]	MS-SSIM, Multi-Scale Structural Similarity Index Measure, is a metric used for image quality assessment that measures the structural similarity between two images across multiple scales.	Document Image Unwarping
26	CLIP-Score↑	[0,1]	The CLIP score measures the similarity between an image and a textual description using the CLIP model, commonly used for image-text matching tasks.	Image Editing
27	FID ↓	[0,∞)	FID, Fr´echet Inception Distance, measures the distance between two multivariate Gaussian distributions: one representing the features of real images and the other representing the features of generated images.	Text-to-Image Generation
28	SAD↓	[0,∞)	SAD, Sum of Absolute Differences, measures the total absolute difference between the predicted and ground truth values for all pixels in an image or region. It is typically used in image matting tasks to assess how closely the model replicates fine-grained image details such as edges and textures.	Object Matting
• Video-related				
29	Fram-Acc↑	[0,1]	Frame-level Accuracy evaluates how many frames (or time steps) in the sequence are correctly classified by comparing predictions with the ground truth on a frame-by-frame basis.	Video Translation
30	FVD ↓	[0,∞)	Fréchet Video Distance (FVD) is a metric used to evaluate the quality of generated video sequences, extending the principles of the FID to videos. FVD calculates the distance between the feature distributions of real and generated videos, taking into account both spatial and temporal dynamics.	Video Generation
31	MUSIQ↑	[0,1]	MUSIQ, Multi-scale Image Quality, quantifies the distance between the feature distributions of real and generated videos.	Video Superresolution
32	absRel ↓	[0,∞)	absRel (Absolute Relative Error) is a metric commonly used in depth estimation tasks, which measures the average ratio of the prediction error to the ground-truth depth.	Video Depth Estimation
33	EPE↓	[0,∞)	End-Point Error (EPE) is a metric commonly used inoptical flowtasks to evaluate the accuracy of predicted motion vectors (flow) between consecutive frames in a video or image sequence. It measures the Euclidean distance between the predicted and ground truth flow vectors at each pixel, providing an average error across the image.	Optical Flow
34	DINO-Score↑	[0, 1]	The DINO Score is calculated as the cosine similarity between the DINOv2 class embedding of two frames, effectively measuring the consistency of a subject's identity across frames. According to the DreamBooth paper, the DINO Score captures more detailed aspects of subject identity compared to the CLIP Score.	Subject-Driven Image Generation
35	L1-Dis↑	[0, 1]	L1-Dis measures the L1 distance between two consecutive frames, evaluating the model's ability to generate static (temporally stable) videos by quantifying the differences between adjacent frames. $\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{T-1}\sum_{t=1}^{T}\frac{\mathcal{L}_1(f_i^t, f_i^{t+1}}{P})$, L1-Dis $= \frac{255-\mathcal{L}_1}{255}$.	Static Video Generation
36	OFS↑	[0, 1]	OFS, Optical Flow Score, captures the movement of pixels between two consecutive frames. By applying a threshold to identify "moving pixels" based on the optical flow, we calculate the ratio of these moving pixels. This ratio quantifies the degree of dynamism in the generated videos.	Dynamic Video Generation

#	Metric	Range	Calculation	Representative Tasks
37	Aesth-Score↑	[0, 1]	The aesthetic score (Aesth-score) is calculated using a ViT with a linear head, as implemented in an improved aesthetic predictor. It has a similar calculation as in CLIP score.	Artistic Content Text-to-Video Generation
38	Suc-Rate↑	[0, 1]	Successful Rate (Suc-Rate) measures the performance of Video Generation. Suc-Rate leverages an open-vocabulary detector to identify the subjects specified in the text prompts. If all the subjects in a text prompt are successfully detected, it classifies the corresponding frame as a successfully generated video frame, and then calculates the ratio of successful frames to the total number of generated frames.	Multi-Class-Conditioned Text-to-Video Generation, Spatial Relation Video Generation, Camera Motion Generation
39	ViCLIP-Score↑	[0, 1]	We calculate the cosine similarity between the ViCLIP embeddings of text prompts and videos. Compared to the CLIP Image model, ViCLIP enables a more comprehensive assessment of the video's style and its overall consistency with the text prompts.	Image-to-Video Generation
40	Avg(DINO+CLIP+OFS+MSS)↑	[0, 1]	For the image-to-video generation task, we conduct a comprehensive evaluation based on Subject Consistency, Background Consistency, Motion Smooth and Dynamic Degree. We utilize the DINO-Score and CLIP-Score to assess subject and background consistency, reflecting the model's ability to adhere to the image prompt. To evaluate the motion smoothness, we measure the Motion Smooth Score (MSS) of the frame-by-frame motion prior via video frame interpolation models. Also, we employ the Optical Flow Score to measure the dynamic degree of the generated videos, ensuring that our metrics do not favor static videos.	Image-to-Video Generation
• 3D-related				
41	AMOTA ↑	[0,1]	AMOTA (Average Multi-Object Tracking Accuracy) is a performance metric used to evaluate multi-object tracking (MOT) systems. It combines detection accuracy and tracking performance by assessing how well a system detects, associates, and tracks multiple objects over time. AMOTA is calculated by averaging tracking accuracy over a range of thresholds for the Intersection over Union (IoU) or matching criteria.	3D Tracking
42	RTE ↓	[0,∞)	Relative Translation Error (RTE) is a metric commonly used in robotics, pose estimation, and SLAM (Simultaneous Localization and Mapping) tasks. It evaluates the accuracy of a system's estimated translation (movement) compared to the ground truth, typically in scenarios where spatial accuracy is crucial.	3D Pose Estimation
43	CD ↓	[0,∞)	Chamfer Distance (CD) is a metric widely used in 3D geometry processing, point cloud generation, and shape matching tasks. It measures the similarity between two sets of points (e.g., two point clouds) by quantifying the average closest-point distance between them. CD is particularly useful for evaluating the alignment and fidelity of reconstructed or generated 3D shapes compared to ground truth data.	Point Cloud Generation
• Segmentation&Detection-related				
44	mIoU↑	[0,1]	mIoU, Mean Intersection over Union, calculates the ratio of the intersection area to the union area between the predicted and ground truth segmentation masks for a single class.	Image Semantic Segmentation
45	m_vIoU↑	[0,1]	m_vIoU measures the spatiotemporal overlap between predicted and ground-truth object regions across multiple video frames	Temporal Action Detection
46	Inst-mIoU↑	[0,1]	Inst-mIoU computes the average IoU score for all part instances across a dataset. It ensures that both over-segmentation and under-segmentation errors are penalized, focusing on instance-level segmentation quality.	3D Part Segmentation
47	PQ↑	[0,1]	PQ, Panoptic Quality, is used for panoptic segmentation tasks, combining both segmentation quality and detection quality. It is a comprehensive metric that evaluates both pixel-level segmentation and object detection quality.	Panoptic Segmentation
48	DICE↑	[0,1]	DICE, Dice Similarity Coefficient, is a metric commonly used in image segmentation tasks, measuring the similarity between the predicted segmentation and the ground truth segmentation.	Bone Fracture Detection
49	S-measure↑	[0,1]	Structure Measure (S-measure) is a metric designed to evaluate the structural similarity between a predicted binary map (e.g., an object mask) and a ground truth binary map. It balances both region-level and boundary-level consistency, ensuring that the evaluation captures holistic structural integrity and fine-grained details.	Video Object Detection
• Audio-related				

#	Metric	Range	Calculation	Representative Tasks
50	CLAP↑	[0,1]	CLAP (Contrastive Language-Audio Pretraining) evaluates the alignment between generated audio and text. It is derived from a contrastive learning framework where embeddings of audio and text are trained to be close in a shared latent space if they are semantically related.	Audio Editing
51	Style-CLAP ↑	[0,1]	Style-CLAP calculates the CLAP cosine similarity between the generated Mel spectrograms and the corresponding textual description of the style to evaluate style fit.	Music Style Transfer
52	MCD ↓	[0,∞)	Mel-cepstral distortion (MCD) measures the spectral distance between the mel-cepstral coefficients (MCCs) of generated speech and reference speech, providing an indication of how closely the generated speech resembles the reference in terms of acoustic characteristics.	Speech Synthesis
53	WER ↓	[0,1]	WER (Word Error Rate) measures the percentage of errors in the transcribed output compared to the reference transcription.	TTS
54	FAD ↓	[0,∞)	Frechet audio distance (FAD) evaluates the quality and realism of generated audio, and measures the similarity between the distribution of features obtained by VGGish in generated audio and those in a set of real (reference) audio samples.	Video-to-Audio
55	PCC ↑	[0,1]	Pitch-Class Consistency (PCC) is a metric used in the evaluation of generated music to assess how consistent the pitch classes (e.g., notes) are across pairs of bars in a piece of music. It measures the overlapping area between the pitch-class histograms of different bars, ensuring that the generated music maintains harmonic coherence.	Music Generation
• Human-aware Evaluation				
56	UPR ↑	[0,1]	UPR, User Preference Rates, UPR measures the proportion of times a particular system or model is preferred over alternatives in a set of user evaluations. It reflects the subjective preferences of users and is often derived from pairwise comparisons or ranking experiments.	Video Style Transfer
57	MOS ↑	[1,5]	Mean Opinion Score (MOS), in which human raters listen to synthesized speech and assess its naturalness, quality, and intelligibility using a 5-point Likert scale.	Speech Generation
58	GPT-Score ↑	[0,1]	GPT-Score evaluates the instruction following rate with GPT assistance, as an alternative to human evaluation.	Audio Question Answering

Mapping Functions of Scoring Metric. Most task evaluation scores, despite utilizing different metrics, fall within a 0-100% range, such as F1, Accuracy (Acc), and ROUGE-L, and follow a monotonically increasing trend. However, certain task metrics produce scores outside this range. For example, regression-related metrics, as well as FID, FVD, and similar metrics, range from 0 to infinity and follow a monotonically decreasing trend. In contrast, MOS scores are represented as a discrete 5-point scale. Due to these varying score ranges across tasks, it becomes intractable to normalize them to a unified scale for level score calculations. Thus, we design the following mapping functions to standardize these metrics into a 1-100% range, thereby streamlining the computation of level scoring algorithms.

- Normalizing **MAE**:

$$y = 2 \times \text{sigmoid}\left(\frac{50}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **RMS**:

$$y = 2 \times \text{sigmoid}\left(\frac{50}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **MSE**:

$$y = 2 \times \text{sigmoid}\left(\frac{5}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **RMSE**:

$$y = 2 \times \text{sigmoid}\left(\frac{5}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **absRel**:

$$y = 2 \times \text{sigmoid}\left(\frac{0.1}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **EPE**:

$$y = 2 \times \text{sigmoid}\left(\frac{1}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **FID**:

$$y = 2 \times \text{sigmoid}\left(\frac{25}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **FVD**:

$$y = 2 \times \text{sigmoid}\left(\frac{100}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **FAD**:

$$y = 2 \times \text{sigmoid}\left(\frac{10}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **PSNR**:

$$y = \tanh\left(\frac{x}{20}\right), \quad \text{where } x \in [0, +\infty), \quad y \in [0, 1).$$

- Normalizing **SAD**:

$$y = 2 \times \text{sigmoid}\left(\frac{10}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **RTE**:

$$y = 2 \times \text{sigmoid}\left(\frac{0.5}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **CD**:

$$y = 2 \times \text{sigmoid}\left(\frac{1}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **MCD**:

$$y = 2 \times \text{sigmoid}\left(\frac{5}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **WER**:

$$y = 1 - x, \quad \text{where } x \in [0, 1], \quad y \in [0, 1].$$

- Normalizing **MS-SSIM**:

$$y = \frac{(x + 1)}{2}, \quad \text{where } x \in [-1, 1], \quad y \in [0, 1].$$

- Normalizing **MOS**:

$$y = \frac{x - 1}{4}, \quad \text{where } x \in [1, 5], \quad y \in [0, 1].$$

B.5 Data Format

To provide a comprehensive understanding of how to utilize our benchmark data, we present examples illustrating how the data files are stored and organized. Figure 8 displays the code snippets showcasing the structures of some representative tasks. Figure 9 illustrates how we organize the benchmark datasets in the file system.

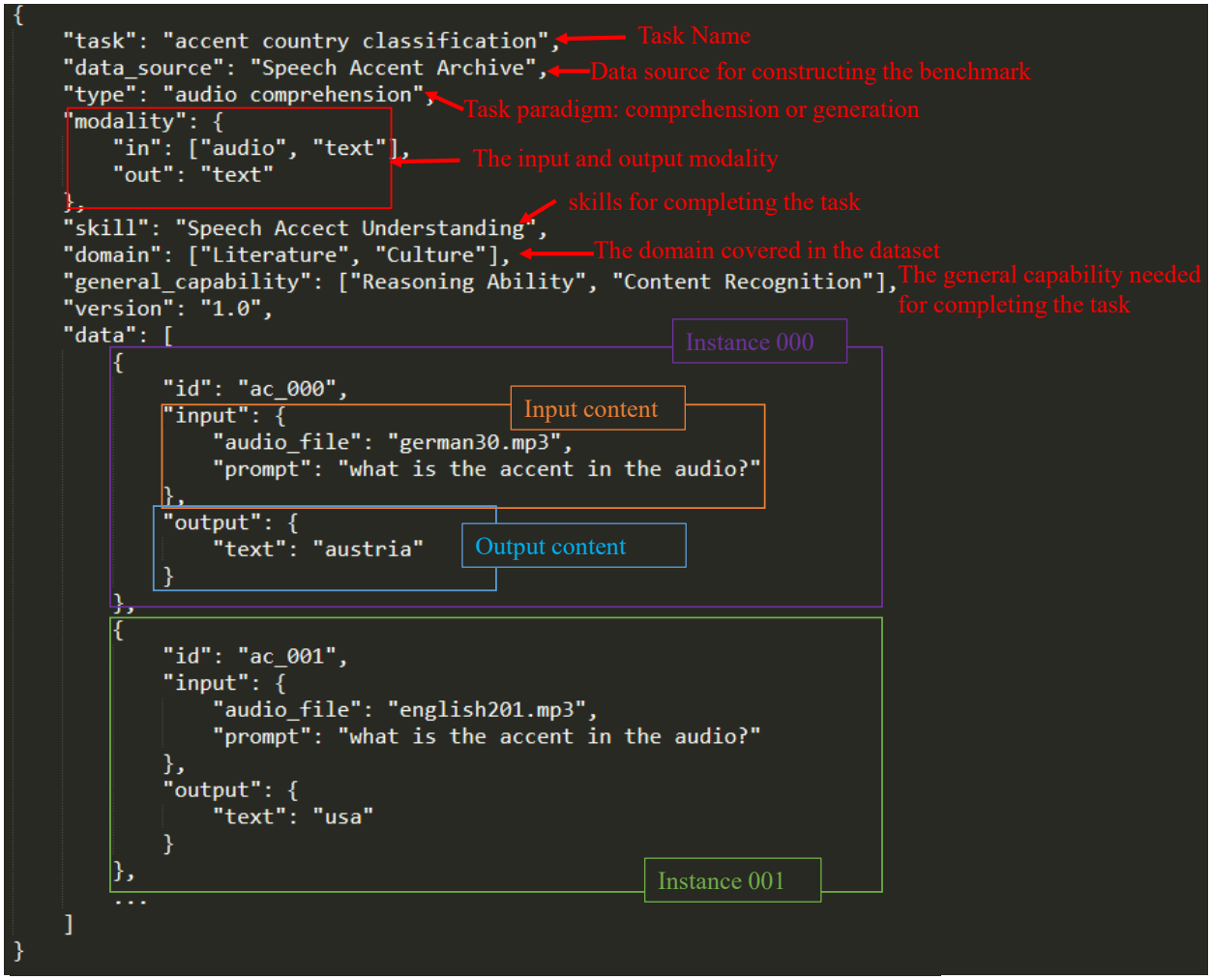


Figure 8: An illustrative example of file formats.

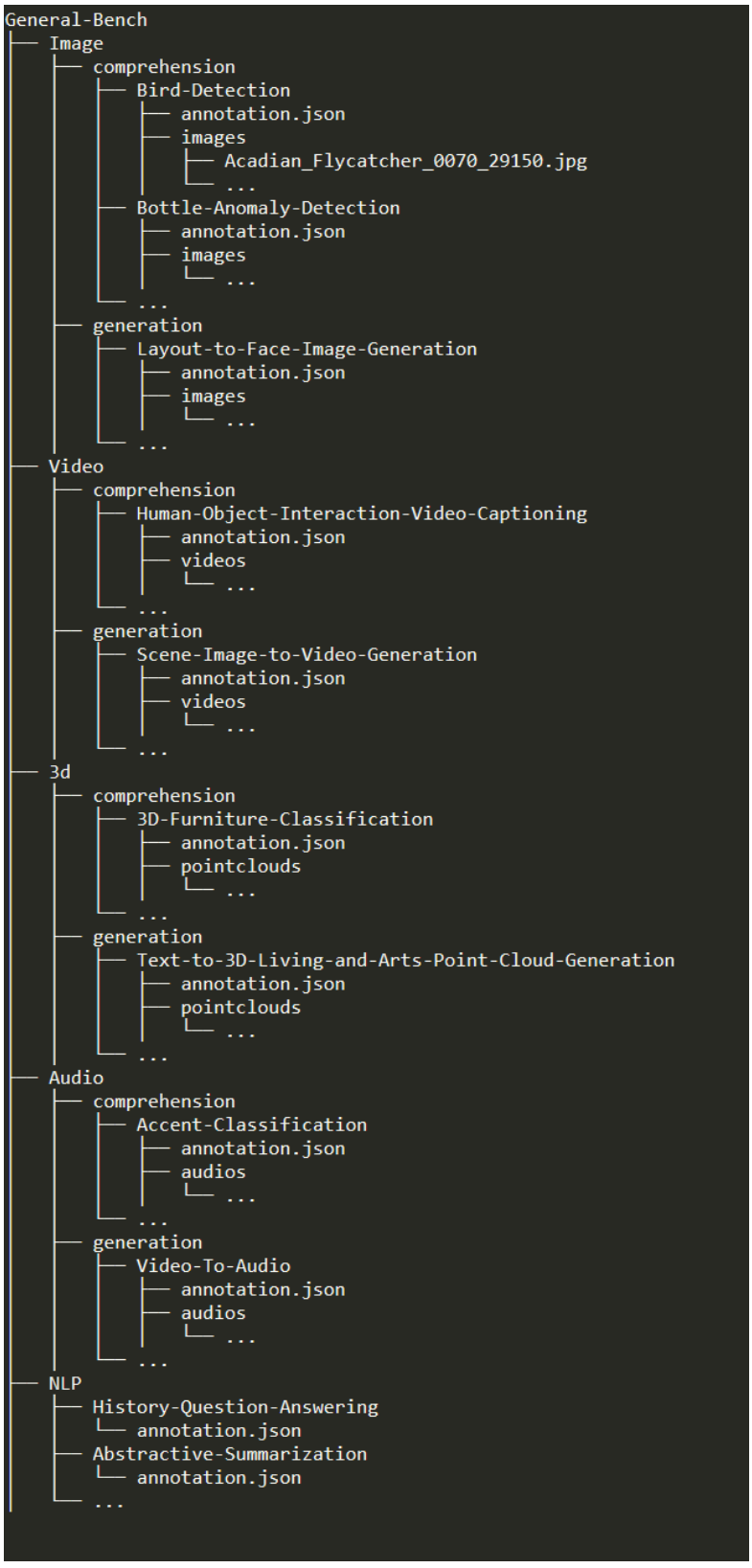


Figure 9: The organization structure of the file system.

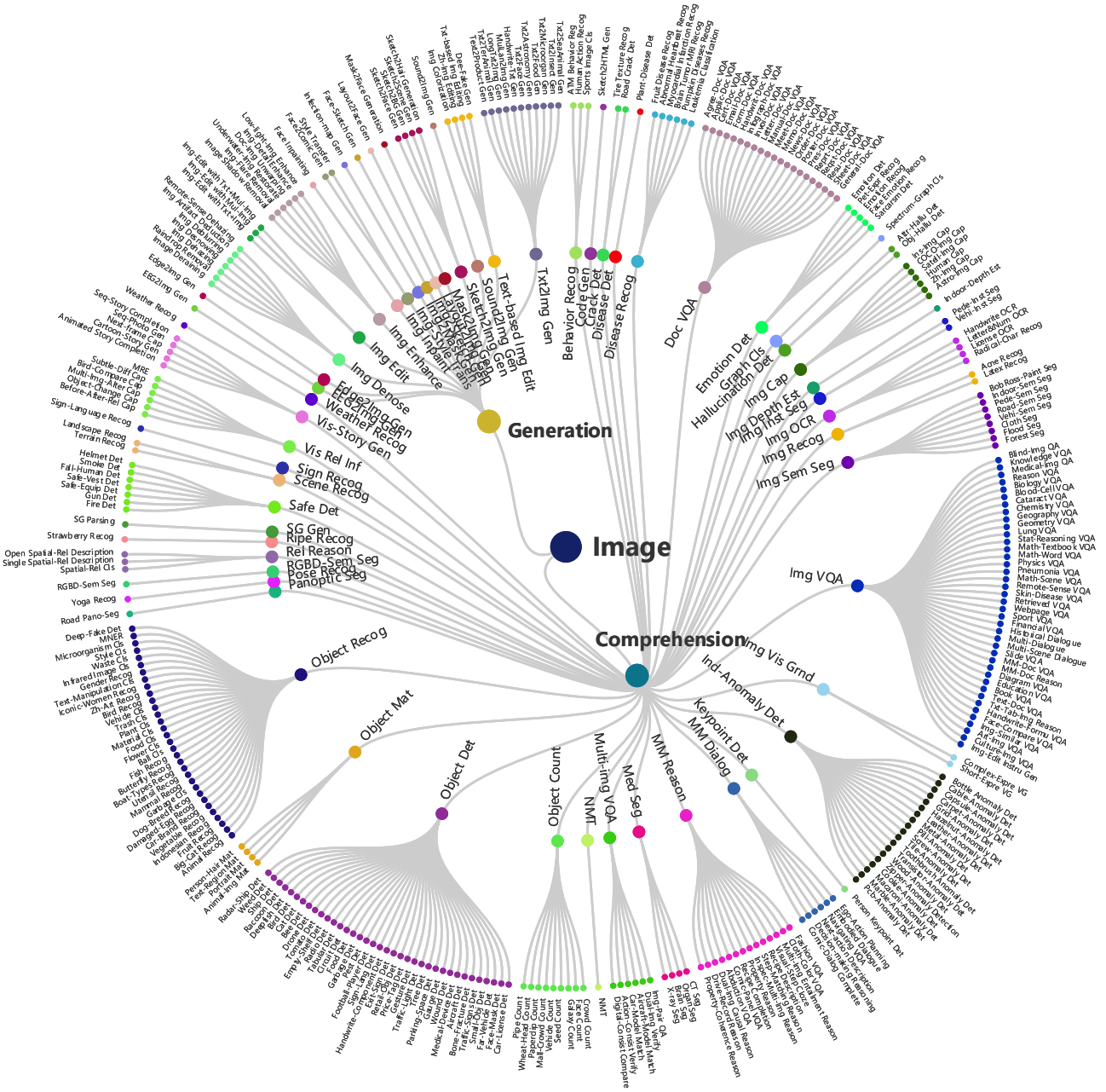


Figure 10: Taxonomy and hierarchy of data in terms of Image modality.

B.6 Data Taxonomy and Hierarchy

We visualize a comprehensive hierarchical taxonomy of our benchmark. Due to space constraints, we have separately illustrated the taxonomy for five major modalities in Figure 10, Figure 11, Figure 12, Figure 13, and Figure 14, respectively. Each visualization includes comprehension and generation paradigms, skills (meta-tasks), and specific tasks for the respective modality.

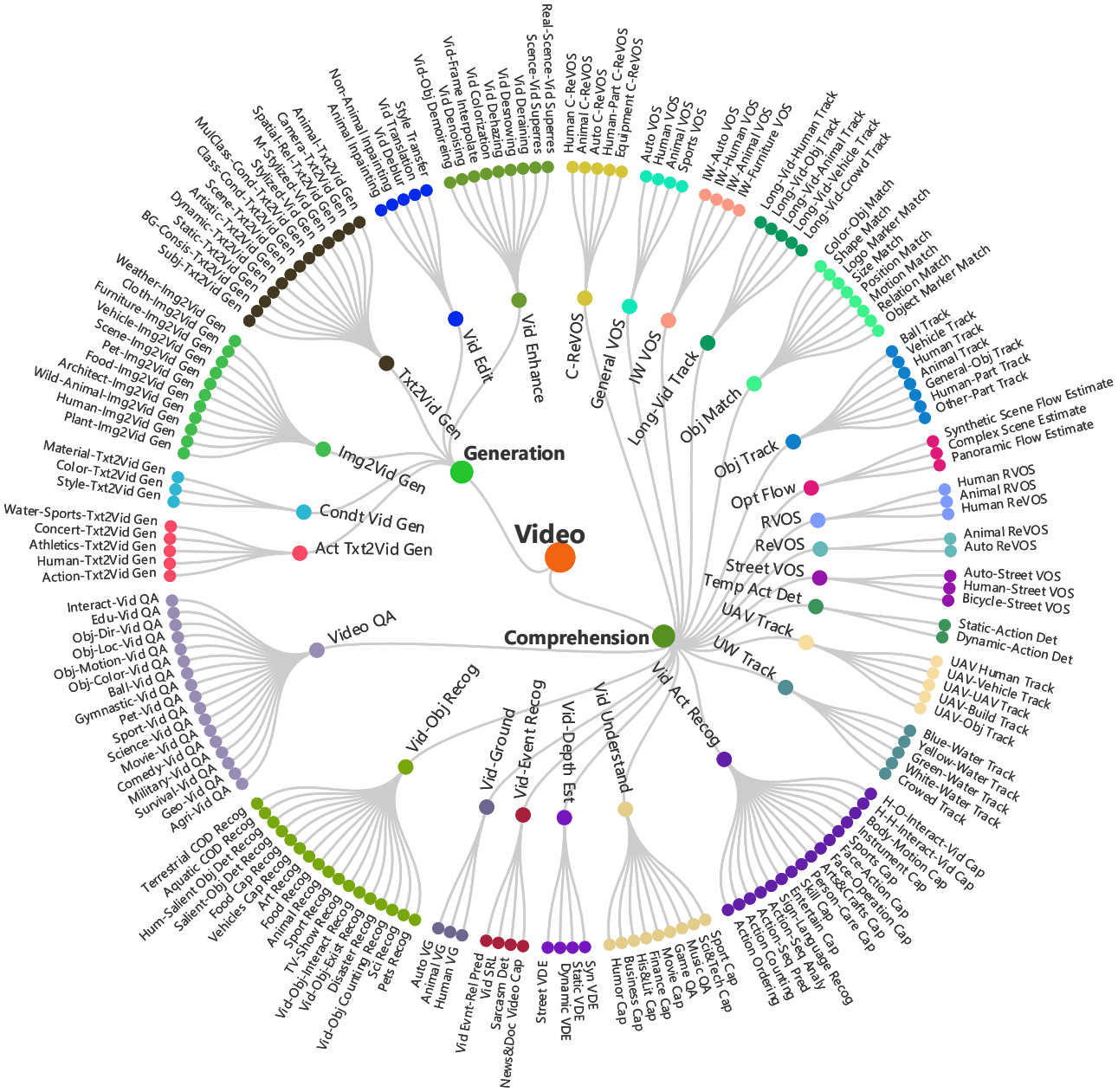


Figure 11: Taxonomy and hierarchy of data in terms of Video modality.

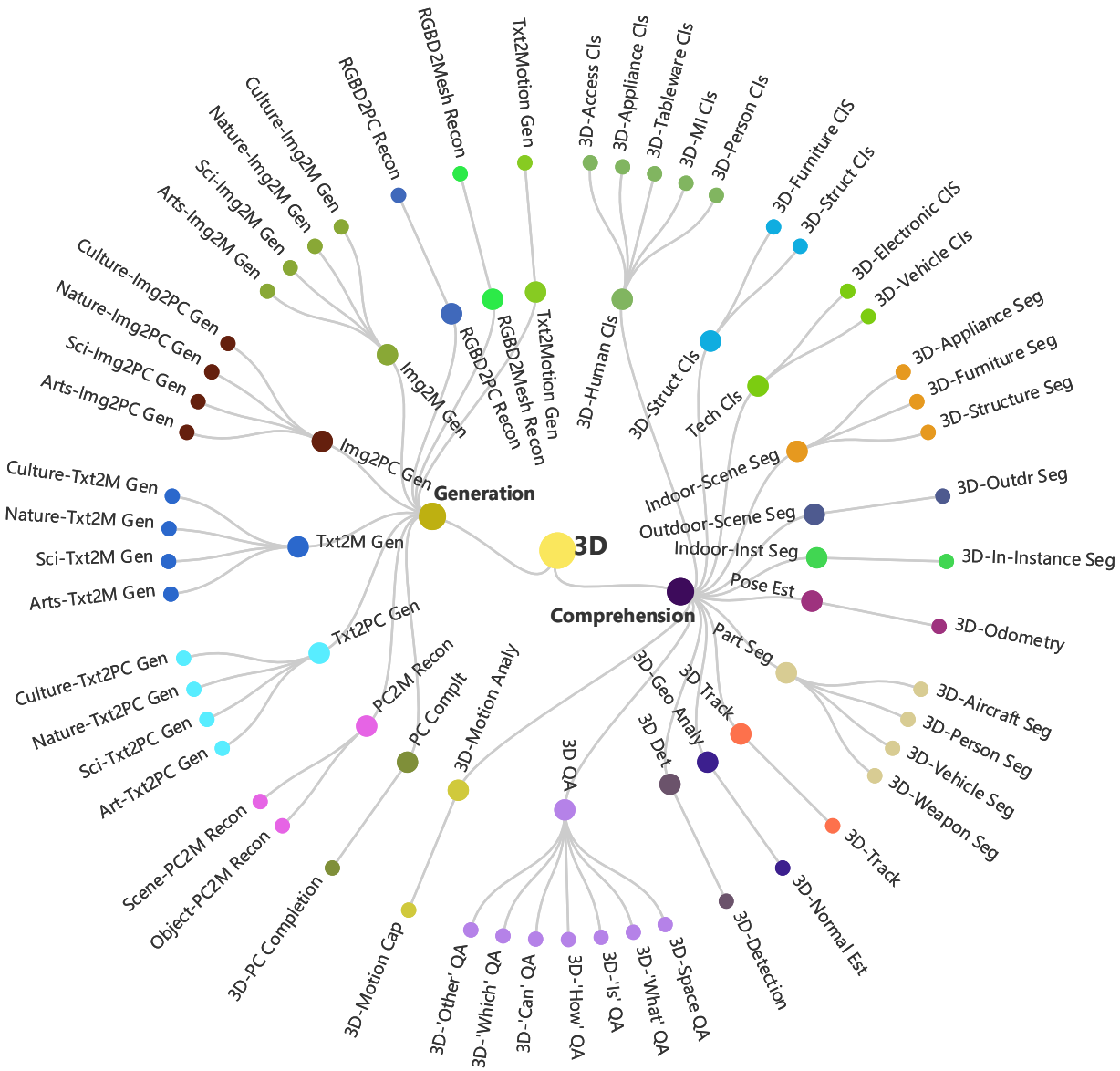


Figure 12: Taxonomy and hierarchy of data in terms of 3D modality.

Figure 13: Taxonomy and hierarchy of data in terms of Audio modality.

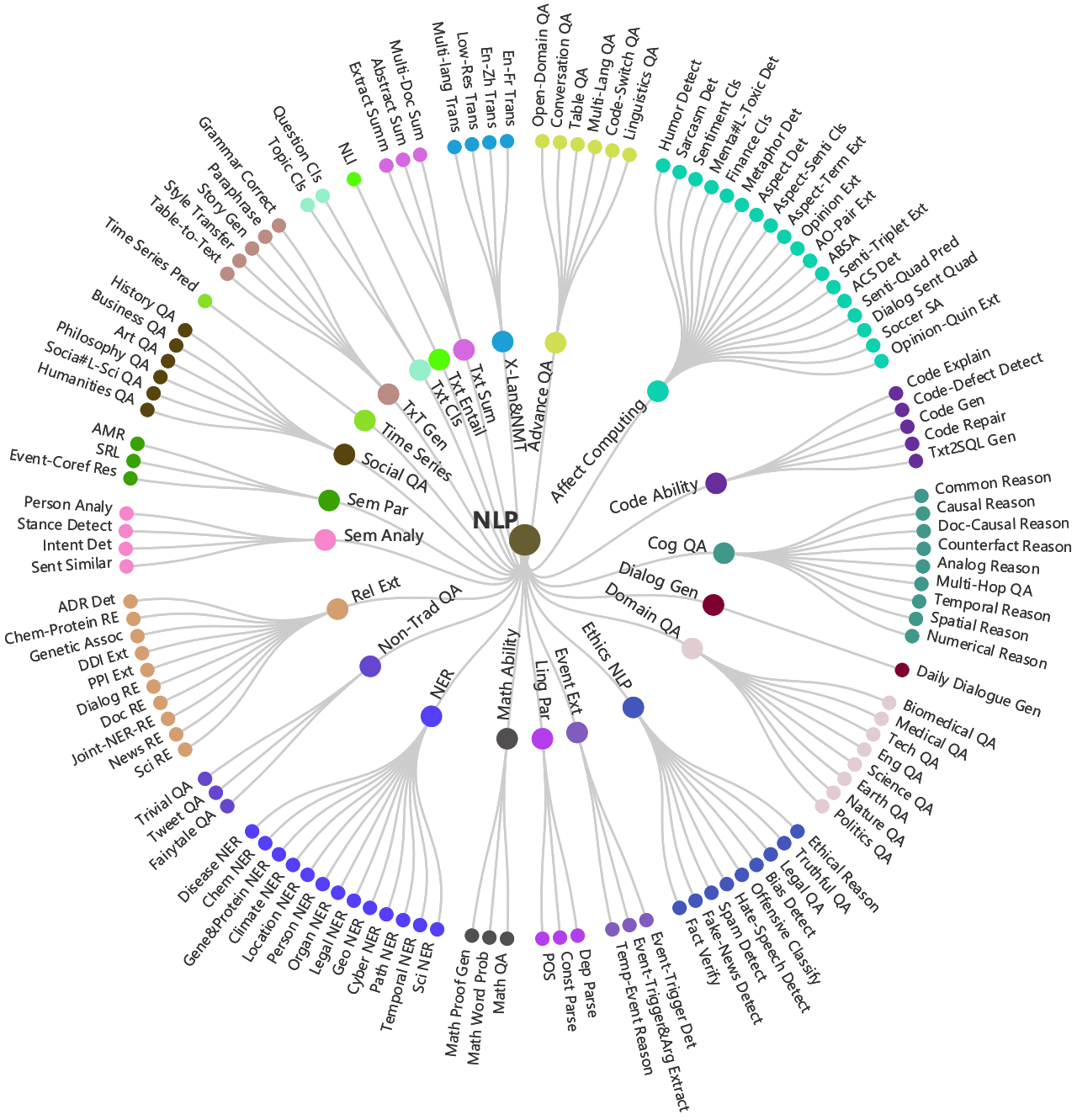


Figure 14: Taxonomy and hierarchy of data in terms of Language modality.

B.7 Data Distributions

Capability. In Figure 15, we present the distribution of capability evaluations across all tasks in General-Bench. These capabilities include: Content Recognition, Commonsense Understanding, Reasoning Ability, Causality Discrimination, Affective Analysis, Problem Solving, Creativity and Innovation, Interactive Capability, and others. As observed, the majority of tasks focus on Content Recognition or perception-related abilities. This emphasis aligns with the current stage of MLLM development, where models are not yet equipped with highly advanced cognitive capabilities. We plan to continuously update the benchmark in the future to accommodate the evolving strengths of more powerful models.

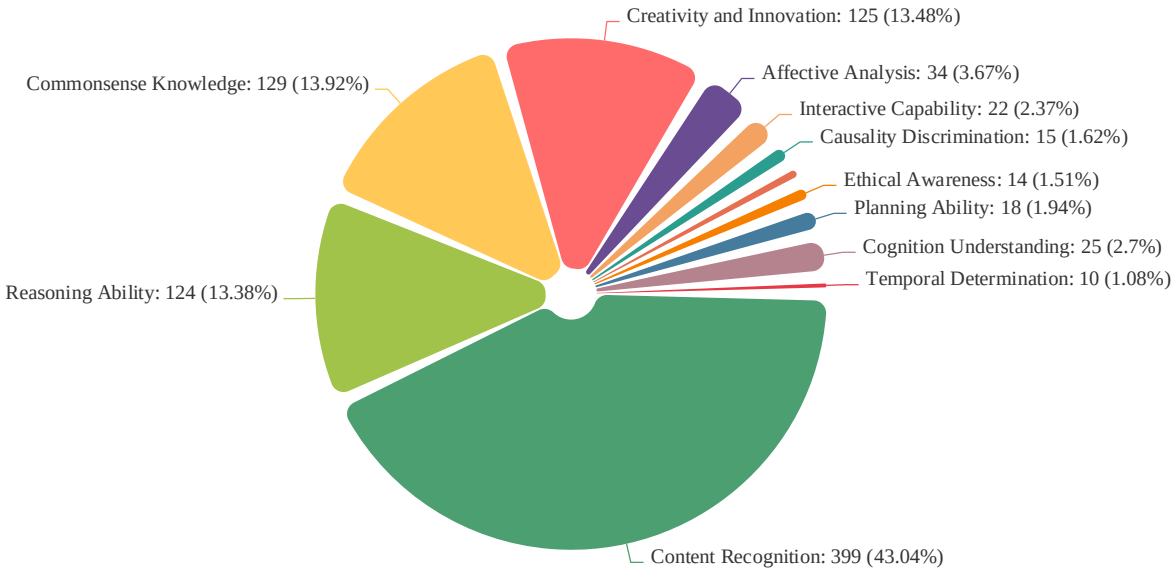


Figure 15: Distribution of various capabilities evaluated in General-Bench.

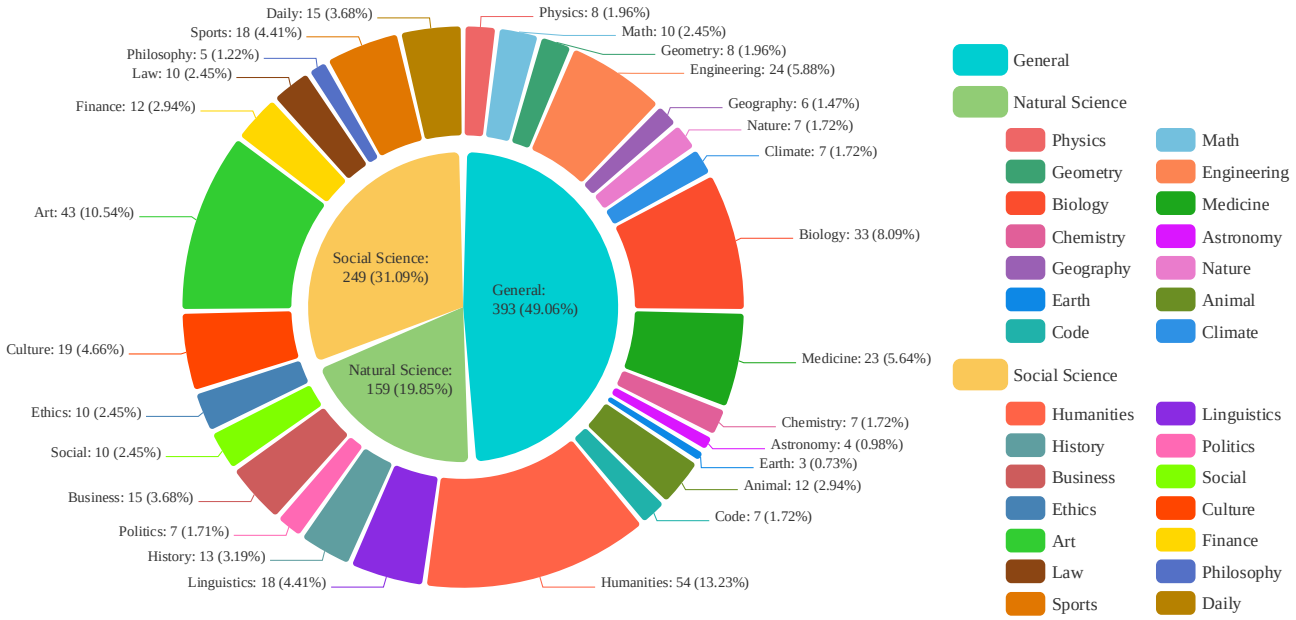


Figure 16: Distribution of various domains and disciplines covered by General-Bench.

Domains and Discipline. Figure 16 illustrates the domains and disciplines covered by our benchmark. While the majority of tasks belong to the general domain, the benchmark also encompasses significant fields from both Physical Sciences and Social Sciences. For Physical Sciences, the benchmark includes disciplines such as Physics, Geometry, Biology, Medicine,

Chemistry, Astronomy, and Geography. For Social Sciences, it spans areas including Humanities, Linguistics, History, Politics, Culture, Art, and Economics. In other words, our benchmark is designed to evaluate the capabilities of MLLMs across a wide range of scientific fields and domains. This ensures the broad evaluative advantage of our benchmark, enabling comprehensive assessment of multimodal generalist models.

Comprehension vs. Generation. We illustrate the task distribution across the two critical paradigms, Comprehension and Generation, in Figure 17. Currently, the majority of tasks are centered on comprehension, which aligns with the present capabilities of most MLLMs.

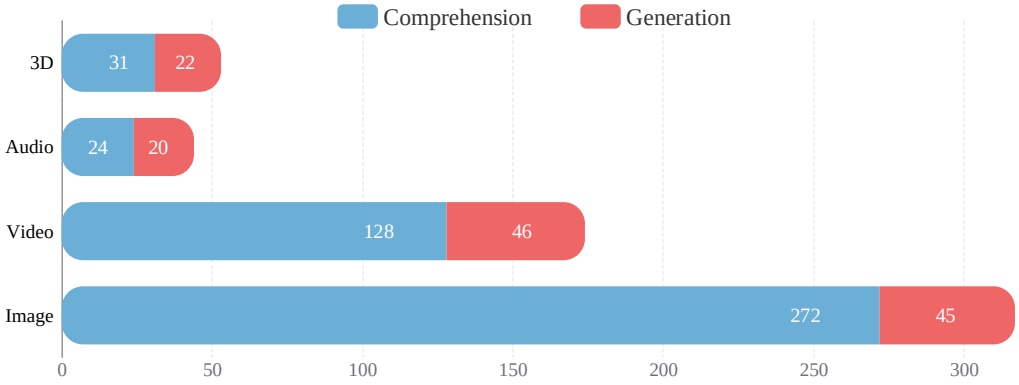


Figure 17: Distribution of various domains and discipline covered by `General-Bench`.

Modality. Finally, we present the distribution of tasks across different modalities in Figure 18. Overall, the image modality constitutes the largest proportion of tasks. We note that beyond the five major modalities—Image, Video, 3D (3D-RGB and Point-Cloud), Audio, and Language—our benchmark also includes tasks in other modalities such as Time Series, Depth, Infrared, Spectrogram, Radar, Code, Document, and Graph. These additional modalities play important roles in specific domains. However, due to the limited number of tasks in these modalities, we have merged and classified their data under broader categories like Image and Language for ease of management.

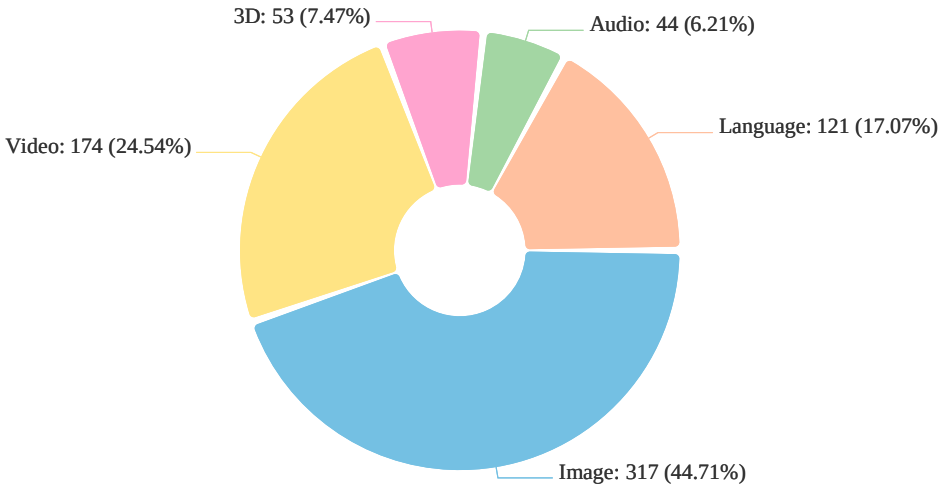


Figure 18: Distribution of different modalities covered in `General-Bench`.

B.8 Extended Data Insights

`General-Bench` particularly places a strong emphasis on the diversity of its evaluation data, covering a wide range of fields and scenarios to assess different aspects of model capabilities, as depicted in Figure 19. First, Table 6 summarizes the statistics of task and skill numbers in `General-Bench`.

Figure 19: **General-Bench** covers over 29 domains, evaluating more than 12 modality-persistent capabilities of generalists, as well as 145 modality-specific skills. In Appendix §B.7 we showcase all tasks and data specification in detail.

B.9 Comparisons with Existing Benchmarks

Table 7 provides a comprehensive comparison of General-Bench with existing MLLM benchmarks. Compared to these benchmarks, General-Bench demonstrates absolute superiority across all comparison aspects. For instance, it is the first MLLM benchmark to cover nearly all commonly used modalities while emphasizing both Comprehension and Generation

Table 6: Summary of numbers of skills, tasks and data instances across modalities.

		Image		Video		Audio		3D		Language	*TOTAL*
		Comp	Gen	Comp	Gen	Comp	Gen	Comp	Gen		
#Skill	Single	40	15	20	6	9	11	13	9	22	145
	Sum	55		26		20		22			
#Task	Single	271	45	126	46	24	20	30	22	118	702
	Sum	316		170		44		52			
#Instance	Single	124,880	26,610	44,442	16,430	11,247	9,516	23,705	10,614	58,432	325,876
	Sum	151,490		60,872		20,763		34,319			

as critical evaluation paradigms. Most notably, `General-Bench` boasts the largest dataset, encompassing 550 tasks with over 275K instances. Also, it evaluates the largest number of MLLMs tested to date, setting a new standard for benchmark comprehensiveness and scale.

Table 7: A full comparison of **`General-Bench`** with other popular MLLM benchmarks.

Aspect / Bench.	Modality	Task Scheme	#Domain	#Skill	#Task	#Sample	Answer Form	#Metric	Annotation	#Tested Models
• Science&Discipline										
ScienceQA (Lu et al., 2022)	Txt,Img	Comp.	12	1	/	21K	MC-QA	Acc.	Repurposed	10
VisAidMath (Ma et al., 2024)	Txt,Img	Comp.	1	1	1	1.2K	MC-QA	Acc.	Repurposed	10
MMMU (Yue et al., 2024)	Txt, Img	Comp.	6	6	30	11.5K	MC-QA	Acc.	Manual	24
NaturalBench (Li et al., 2024e)	Txt,Img	Comp.	1	/	27	7.6K	MC-QA	Acc.	Repurposed	32
• Audio										
AudioBench (Wang et al., 2024b)	Txt, Aud	Comp.	/	3	8	100K	Free-Form	WER,METEOR Llama-score	Repurposed	4
MARBLE (Yuan et al., 2023b)	Txt, Aud	Comp.	/	12	18	/	Free-Form	Origin(19)	Repurposed	0
MMAU (Sakshi et al., 2024)	Txt, Aud	Comp.	/	12	27	10K	MC-QA	Acc.	Repurposed Manual	14
• 3D										
T^{3}Bench (He et al., 2023)	Txt, 3D	Gen.	/	3	3	300	Free-From	Origin(2)	Repurposed	0
3DBench (Zhang et al., 2024b)	Txt, 3D	Comp.	/	12	12	8K	Free-From	Origin(6)	Repurposed	3
• Video										
Video-Bench (Ning et al., 2023)	Txt,Vid	Comp.	/	3	7	17K	MC-QA	Acc.	Repurposed	8
VideoMME (Fu et al., 2024b)	Txt,Vid	Comp.	6	12	12	900	MC-QA	Acc.	Manual	13
MLVU (Zhou et al., 2024a)	Txt, Vid	Comp.	7	9	9	2593	MC-QA Open	Acc. GPT-Ranking	Manual	20
MVBench (Li et al., 2024f)	Txt,Vid	Comp.	/	20	20	4k	MC-QA	Acc.	Repurposed	14
AutoEval-Video (Chen et al., 2024d)	Txt,Vid	Comp.	12	9	9	327	Open	GPT-score	Manual	11
VBench (Huang et al., 2024)	Txt,Vid	Gen.	8	16	16	100	Open	/	Manual	4
• Image										
MME (Fu et al., 2023)	Txt,Img	Comp.	/	14	14	2.2K	MC-QA	Acc.	Repurposed	30

Aspect / Bench.	Modality	Task Scheme	#Domain	#Skill	#Task	#Sample	Answer Form	#Metric	Annotation	#Tested Model
LVLM-eHub (Xu et al., 2023b)	Txt, Img	Comp.	/	6	47	2.1K	MC-QA Open	Acc. CIDEr top-1 Acc.	Repurposed	8
MM-Vet (Yu et al., 2024)	Txt,Img	Comp.	/	6	1	205	MC-QA	GPT-score	Repurposed	16
V*Bench (Wu and Xie, 2024)	Txt,Img	Comp.	1	2	2	191	MC-QA	Acc.	Manual	13
MMIE (Xia et al., 2024b)	Txt, Img	Comp. Gen.	10	4	4	20K	MC-QA Open	/	Manual	8
Mia-bench (Qian et al., 2024)	Txt, Img	Comp.	15	8	8	400	Open	GPT-score	Manual	29
MME-RealWorld (Zhang et al., 2024c)	Txt, Img	Comp.	5	43	43	29.4K	MC-QA	Acc.	Manual	29
MLLM-Bench (Ge et al., 2024)	Txt, Img	Comp.	/	6	6	420	MC-QA Open	GPT-score	Manual	21
Q-Bench (Wu et al., 2024b)	Txt, Img	Comp.	1	3	3	84.7K	MC-QA Open	GPT-score	Manual+ Repurposed	15
MUIRBench (Wang et al., 2024c)	Txt,Img	Comp.	12	12	12	2.6K	MC-QA	Acc.	Manual+ Repurposed	20
MileBench (Song et al., 2024)	Txt,Img	Comp.	/	2	2	6.4K	MC-QA Open	Acc. ROUGE-L	Manual+ Repurposed	22
MMBench (Liu et al., 2024a)	Txt,Img	Comp.	/	2	20	3K	MC-QA	Acc.	Repurposed	21
• Omini										
SEED-Bench (Li et al., 2023b)	Txt,Img,Vid	Comp.	/	12	12	19K	MC-QA	Acc.	Manual	18
SEED-Bench-2 (Li et al., 2023c)	Txt,Img,Vid	Comp. Gen.	3	22	22	24K	MC-QA	Acc.	Manual+ Repurposed	23
CV-Bench (Tong et al., 2024a)	Txt, Img, 3D	Comp.	2	4	4	2.6K	MC-QA	Acc.	Repurposed	15
MMIU (Meng et al., 2024b)	Txt,Img,Vid, Point-Cloud,Depth	Comp.	/	7	52	11.7K	MC-QA	Acc.	Repurposed	22
MMT-Bench (Ying et al., 2024b)	Txt,Img,Vid, Point-Cloud	Comp.	/	32	162	31K	MC-QA	Acc.	Repurposed	30
MEGA-Bench (Chen et al., 2024e)	Txt,Img,Vid	Comp.	/	10	505	8K	Free-Form	Origin (45)	Manual	22
General-Bench (Ours)	Txt,Img,Vid,Aud, Time,Depth,3D-RGB, Point-Cloud,Infrared, Spectrogram,Radar, Code,Doc,Graph,· · ·	Comp.+Gen.	**29**	**145**	**702**	**325.8K**	Free-Form	Origin (**58**)	Reannotated +Manual	**172 × Specialists & 102 × Generalists**

C Extension on Experimental Results

In the main text, we presented a subset of the evaluation performance of MLLMs at the skill level due to space constraints. Here, we provide the complete results on all specific tasks. The results are organized based on modality and task paradigm distinctions.

It is important to note that we conducted inference using different MLLMs' open-source codebases or APIs. However, the choice of prompts for different tasks can significantly impact the model's performance on a given task. For instance, some models require highly detailed and specific in-context instructions in the input prompt to achieve their best performance. To ensure fairness, our team applied a uniform prompt across all MLLMs for a given task, without any model-specific prompt tuning. If the model developers are unsatisfied with the presented results, we welcome them to adjust the prompts to obtain more representative and improved scores.

C.1 Multimodal Specialist and Generalist Systems

SoTA Specialist. For each specific task under a specific modality, we select a SoTA specialist to generate benchmark results. The selection of specialists is determined based on two criteria: 1) their performance on each task using public benchmarks and leaderboards, i.e., they must demonstrate top performance; and 2) whether they are widely recognized and utilized by the community. Meanwhile, we exclude models that lack reliable open-source code or parameters (as we are unable to run our own data through them), even if such models claim to be SoTA in their own papers. It is important to note that the specialists we use must have undergone large-scale supervised pretraining on the corresponding tasks, enabling them to achieve SoTA performances. In our implementation, we directly load their released parameters and perform inference on the `General-Bench` test sets. In total, we have 172 specialists.

Multimodal Generalists. We consider a diverse set of existing popular MLLMs that are capable of handling specific or various modalities and tasks. This includes both open-source systems and closed-source ones (such as the OpenAI GPT series). For open-source models, we implement them by loading their released parameters and directly performing inference on the `General-Bench` test sets. For closed-source models, we utilize their APIs to access the services. We note that, despite the release of a vast number of MLLMs in the community, due to resource constraints, we only consider a subset of MLLMs that demonstrate strong and stable capabilities and are widely recognized and utilized. However, our evaluation system remains open, and we encourage more MLLMs interested in our benchmarking system to participate by running their own evaluations and submitting their scores. Table 9 summarizes all the multimodal generalists employed, including their corresponding modality support, characterized skills, parameter sizes, and backbone LLM architectures.

Table 9: A complete list of (multimodal) generalists evaluated on General-Bench.

#	Model	Backbone	Size	Modality Support	Paradigm
• Language-oriented (Closed/Open-sourced) Models					
1	Meta-Llama-3.1-8B-Instruct (Touvron et al., 2023)	Llama	8B	Language	/
2	Gemma-2-9b-it (Team et al., 2024b)	Gemma	9B	Language	/
3	GPT-J (Wang and Komatsuzaki, 2021)	GPT-J	6B	Language	/
4	ChatGLM-6B (GLM et al., 2024)	ChatGLM	6B	Language	/
5	Qwen2.5-7B-Instruct (Yang et al., 2024)	Qwen2.5	7B	Language	/
6	InternLM2-Chat-7B (Cai et al., 2024)	InternLM2	7B	Language	/
7	Baichuan2-7B-Chat (Yang et al., 2023)	Baichuan2	7B	Language	/
8	Vicuna-7b-V1.5 (Chiang et al., 2023)	Vicuna	7B	Language	/
9	Falcon3-7B-Instruct (Almazrouei et al., 2023)	Falcon3	7B	Language	/

#	Model	Backbone	Size	Modality Support	Paradigm
10	Ministral-8B-Instruct-2410 (Jiang et al., 2024a)	Ministral	8B	Language	/
11	Yi-lightning (Young et al., 2024)	Llama	6B	Language	/
12	GPT-3.5-turbo (OpenAI, 2022a)	GPT3.5	/	Language	/
• Multimodal Close-sourced Models					
1	GPT4-V (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
2	GPT4-o-mini (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
3	GPT4-o (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
4	GPT4-o-4096 (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
5	ChatGPT-o-latest (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
6	Claude-3.5-Sonnet (Team, 2024)	Claude-3.5-Sonnet	/	Language, Image	Comprehension
7	Claude-3.5-Opus (Team, 2024)	Claude-3.5-Opus	/	Language, Image	Comprehension
8	Gemini-1.5-Pro (Team et al., 2024a)	Gemini	/	Language, Image	Comprehension
9	Gemini-1.5-Flash (Team et al., 2024a)	Gemini	/	Language, Image	Comprehension
• Multimodal Open-sourced Models					
1	Yi-vision-v2 (Young et al., 2024)	LLaVa	6B	Language, Image	Comprehension
2	Emu2-37B (Sun et al., 2024)	LLaMA-33B	37B	Language, Image	Comprehension+Generation
3	InternVL2.5-2B (Chen et al., 2024c)	internlm2_5-1_8b-chat	2B	Language, Image	Comprehension
4	InternVL2.5-4B (Chen et al., 2024c)	Qwen2.5-3B-Instruct	4B	Language, Image	Comprehension
5	InternVL2.5-8B (Chen et al., 2024c)	internlm2_5-7b-chat	8B	Language, Image	Comprehension
6	Mini-InternVL-Chat-2B-V1-5 (Gao et al., 2024)	InternLM2-Chat-1.8B	2B	Language, Image	Comprehension
7	Mini-InternVL-Chat-4B-V1-5 (Gao et al., 2024)	Phi-3-mini-128k-instruct	4B	Language, Image	Comprehension
8	InternLM-XComposer2-VL-1.8B (Dong et al., 2024)	InternLM2-Chat-1.8B	1.8B	Language, Image	Comprehension
9	MoE-LLAVA-Phi2-2.7B-4e-384 (Lin et al., 2024)	Phi2	2.7B	Language, Image	Comprehension
10	Monkey-10B-chat (Li et al., 2024g)	Qwev-7B	10B	Language, Image	Comprehension
11	mPLUG-Owl2-LLaMA2-7b (Ye et al., 2024)	LLaMA2-7b	7B	Language, Image	Comprehension
12	Phi-3.5-Vision-Instruct (Abdin et al., 2024)	Phi-3 Mini	4.2B	Language, Image	Comprehension
13	Cambrian-1-8B (Tong et al., 2024b)	LLaMA3-8B-Instruct	8B	Language, Image	Comprehension
14	DetGPT (Pi et al., 2023)	Vicuna-7B	7B	Language, Image	Comprehension
15	Otter (Li et al., 2023d)	LLaMA-7B	7B	Language, image	Comprehension
16	NExT-Chat (Zhang et al., 2023c)	LLaVA	7B	Language, Image	Comprehension
17	GPT4RoI-7B (Zhang et al., 2023d)	LLaMA-7B	7B	Language, Image	Comprehension
18	GLaMM (Rasheed et al., 2024)	Vicuna-7B	7B	Language, Image	Comprehension

#	Model	Backbone	Size	Modality Support	Paradigm
19	Pixtral-12B (Agrawal et al., 2024)	Mistral-Nemo-12B	12B	Language, Image	Comprehension
20	BLIP-2 (Li et al., 2023a)	Flan T5-xl	3B	Language, Image	Comprehension
21	BLIP-3 (XGen-MM) (Xue et al., 2024)	Phi3-mini	4B	Language, Image	Comprehension
22	miniMonkey (Li et al., 2024g)	Qwev-7B	7B	Language, Image	Comprehension
23	MiniGPT4-LLaMA2-7B (Zhu et al., 2023a)	LLaMA2-7B-instruct	7B	Language, Image	Comprehension
24	Show-o (Xie et al., 2024)	Show-o	1.3B	Language, Image	Comprehension+Generation
25	DeepSeek-VL-7B-Base (Lu et al., 2024b)	DeepSeek-LLM-7b-base	7B	Language, Image	Comprehension
26	DeepSeek-VL-7B-Chat (Lu et al., 2024b)	DeepSeek	7B	Language, Image	Comprehension
27	LISA (Lai et al., 2024)	LLaMA-7B	7B	Language, Image	Comprehension
28	CogVLM-Chat (Wang et al., 2023b)	Vicuna-v1.5-7B	17B	Language, Image	Comprehension
29	ShareGPT4V-7B (Chen et al., 2025)	Vicuna-v1.5-7B	7B	Language, Image	Comprehension
30	ShareGPT4V-13B (Chen et al., 2025)	Vicuna-v1.5-13B	13B	Language, Image	Comprehension
31	GLM-VL-Chat (Du et al., 2021)	GLM-4V	9B	Language, Image	Comprehension
32	OMG-LLaVA-InternLM20B (Zhang et al., 2024a)	internlm2-7b	7B	Language, Image	Comprehension
33	Idefics3-8B-Llama3 (Laurençon et al., 2024)	Llama-3.1-8B	8B	Language, Image	Comprehension
34	MiniCPM3-4B (Hu et al., 2024a)	MiniCPM3-4B	4B	Language, Image	Comprehension
35	SEED-LLaMA-13B (Ge et al., 2023)	Llama2-chat-13B	14B	Language, Image	Comprehension+Generation
36	LaVIT-V2 (7B) (Jin et al.)	LLaMA-7B	7B	Language, Image	Comprehension+Generation
37	LM4LV (Zheng et al., 2024)	LLaMA2-7B instruct	7B	Language, Video	Generation
38	CoLVA-2B (Zhou et al., 2025)	Qwen2-2B	2B	Language, Image, Video	Comprehension
39	CoLVA-4B (Zhou et al., 2025)	Phi3-3.8B	4.1B	Language, Image, Video	Comprehension
40	Long-LLaVA-9B (Wang et al., 2024d)	Jamba-9B-Instruct	9B	Language, Video	Comprehension
41	DeepSeek-VL-2-small (Lu et al., 2024b)	DeepSeekMoE-16B	2.8B	Language, Image	Comprehension
42	DeepSeek-VL-2 (Lu et al., 2024b)	DeepSeekMoE-27B	4.5B	Language, Image	Comprehension
43	Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B	7B	Language, Image, Video	Comprehension
44	Qwen-Audio-Chat (Chu et al., 2023)	Qwen-7B	7B	Language, Audio	Comprehension
45	Qwen2-VL-7B (Wang et al., 2024a)	Qwen2-7B	7B	Language, Image, Video	Comprehension
46	Qwen2-Audio-Instruct (Chu et al., 2024)	Qwen-7B	7B	Language, Audio	Comprehension
47	Qwen2-VL-72B (Wang et al., 2024a)	Qwen2-72B	72B	Language, Image, Video	Comprehension
48	LLaVA-NeXT-13B (Liu et al., 2024b)	Vicuna-13B	13B	Language, Image	Comprehension
49	LLaVA-NeXT-34B (Liu et al., 2024b)	Nous-Hermes-2-Yi-34B	34B	Language, Image	Comprehension

#	Model	Backbone	Size	Modality Support	Paradigm
50	LLaVA-One-Vision-7B (Li et al., 2024d)	Qwen2-7B	7B	Language, Image, Video	Comprehension
51	LLaVA-One-Vision-72B (Li et al., 2024d)	Qwen2-72B	72B	Language, Image, Video	Comprehension
52	Sa2VA-8B (Yuan et al., 2025)	InternLM2-7B	8B	Language, Image, Video	Comprehension
53	Sa2VA-26B (Yuan et al., 2025)	InternLM2-20B	26B	Language, Image, Video	Comprehension
54	InternVL-2-8B (Chen et al., 2024c)	InternLM2-7B	8B	Language, Image, Video	Comprehension
55	InternVL-2.5-8B (Chen et al., 2024c)	internlm2_5-7b-chat	8B	Language, Image, Video	Comprehension
56	InternVL-2-26B (Chen et al., 2024c)	InternLM2-20B	26B	Language, Image, Video	Comprehension
57	InternVL-2.5-26B (Chen et al., 2024c)	internlm2_5-20b-chat	26B	Language, Image, Video	Comprehension
58	Vitron-V1 (Fei et al., 2024a)	vicuna-7b-v0	7B	Language, Image, Video	Comprehension+Generation
59	Mini-Gemini (Li et al., 2024c)	Nous-Hermes-2-Yi-34B	34B	Language, Image	Comprehension+Generation
60	3D-LLM-2.1B (Hong et al., 2023)	BLIP2	2.1B	Language, 3D	Comprehension
61	PointLLM-7B (Xu et al., 2025)	LLaMA	7B	Language, 3D	Comrehension
62	PointLLM-13B (Xu et al., 2025)	LLaMA	13B	Language, 3D	Comprehension
63	3D-VisTA (Zhu et al., 2023b)	BERT	1.3B	Language, 3D	Comprehension
64	AvatarGPT (Zhou et al., 2024b)	T5-large	770M	Language, 3D	Comprehension
65	MotionGPT-T5 (Jiang et al., 2024b)	T5	220M	Language, 3D	Generation
66	MotionGPT-LLaMA (Zhang et al., 2023e)	LLaMA	13B	Language, 3D	Generation
67	LLaMA-mesh (Zhang et al., 2023e)	LLaMA	7B	Language, 3D	Generation
68	GAMA (Ghosh et al., 2024)	Llama-2-7b-chat	7B	Language, Audio	Comprehension
69	Pengi (Deshmukh et al., 2023)	GPT2-base	124M	Language, Audio	Comprehension
70	WavLLM (Hu et al., 2024b)	LLaMA-2-7B-chat	7B	Language, Audio	Comprehension
71	SALMONN-7B (Tang et al., 2023)	Vicuna-7B	7B	Language, Audio (Speech)	Comprehension
72	SALMONN-13B (Tang et al., 2023)	Vicuna-13B	13B	Language, Audio (Speech)	Comprehension
73	SpeechGPT-7B-com (Zhang et al., 2023a)	LLaMA-2	7B	Language, Audio (Speech)	Generation
74	AudioGPT-GPT4 (Huang et al., 2023)	GPT-4	/	Language, Audio (Speech, Sound)	Generation
75	AnyGPT (Zhan et al., 2024)	LLaMA-2-7B	8B	Language, Image, Audio (Speech, Music)	Comprehension+Generation
76	PandaGPT-13B (Su et al., 2023)	Vicuna-13B-v0	13B	Language, Image, Video, Audio	Comprehension
77	ImageBind-LLM (Han et al., 2023)	LLama-1-7B	7B	Language, Image, Video, Audio	Comprehension
78	ModaVerse-7b-v0 (Wang et al., 2024e)	Vicuna-7b-V0	7B	Language, Image, Video, Audio	Comprehension+Generation
79	Unified-io-2-XXL (Lu et al., 2024a)	UIO-2-XXL	6.8B	Language, Image, Video, Audio	Comprehension+Generation

Table 11: Performance of multimodal generalists on video comprehension and generation skills.

Model	Video Comprehension Skill (Avg within each #V-C Group) #1 #11	#2 #12	#3 #13	#4 #14	#5 #15	#6 #16	#7 #17	#8 #18	#9 #19	#10 #20	Task Completion #Supported Task	#Win-over-Specialist	Level Score on Video Level-2	Level-3	Level-4
SoTA Specialist	37.43 45.84	49.64 13.92	21.31 0.14	23.06 48.06	81.85 68.96	85.43 63.62	54.53 77.02	64.83 75.08	40.65 37.20	30.80 44.00	/	/	/	/	/
InternVL-2.5-8B	33.15 0.00	27.54 0.00	14.51 0.00	18.83 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 4.85	55 (43.7%)	5 (4.0%)	5.76	1.24	0.00
InternVL-2.5-26B	37.03 0.00	32.01 0.00	18.71 0.00	21.57 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 5.30	55 (43.7%)	26 (20.6%)	6.70	3.76	0.00
Qwen2-VL-72B	38.22 0.00	32.32 0.00	19.35 0.00	22.70 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 5.70	55 (43.7%)	22 (17.5%)	6.89	5.22	0.00
DeepSeek-VL-2	21.50 0.00	18.90 0.00	12.10 0.00	12.10 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 3.20	55 (43.7%)	5 (4.0%)	3.98	0.64	0.00
LLaVA-One-Vision-72B	31.20 0.00	31.30 0.00	19.10 0.00	10.60 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 1.70	56 (44.4%)	21 (16.7%)	5.83	3.75	0.00
Sa2VA-8B	33.19 0.00	25.11 60.28	16.75 0.00	8.67 0.00	0.00 19.85	0.00 37.83	0.00 46.36	71.03 42.58	50.95 48.02	0.00 1.48	91 (72.2%)	32 (25.4%)	8.31	4.38	0.00
Sa2VA-26B	35.33 0.00	26.33 0.00	17.58 0.00	10.39 0.00	0.00 28.41	0.00 38.91	0.00 47.10	0.00 43.12	0.00 48.42	0.00 1.70	81 (64.3%)	27 (21.4%)	8.81	4.58	0.00
CoLVA-4B	32.68 0.00	26.45 0.00	13.55 0.00	17.62 0.00	0.00 45.81	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 4.23	63 (50.0%)	8 (6.3%)	4.78	1.24	0.00
InternVL-2-8B	32.69 0.00	27.09 0.00	14.24 0.00	17.61 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 4.85	55 (43.7%)	0 (0.0%)	5.64	0.46	0.00
Long-LLaVA-9B	36.14 0.00	26.25 0.00	15.89 0.00	15.53 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 4.20	54 (42.9%)	22 (17.5%)	5.84	3.81	0.00

Model	Video Generation Skill (Avg within each #V-G Group) #1	#2	#3	#4	#5	#6	Task Completion #Task-Supprt	#Win-Spclst	Level Score on Video Level-2	Level-3	Level-4
SoTA Specialist	69.09	55.79	88.94	62.90	37.79	51.46	/	/	/	/	/
VidAgent	52.42	47.73	88.84	63.61	0.00	0.00	30 (65.2%)	0 (0.0%)	25.00	0.00	0.00
LM4LV	0.00	0.00	0.00	0.00	25.90	5.93	8 (17.4%)	0 (0.0%)	6.74	0.00	0.00
NExT-GPT-V1.5	26.78	6.72	130.22	16.03	0.08	0.06	40 (87.0%)	0 (0.0%)	8.34	0.71	0.00
Vitron-V1	36.74	19.32	116.31	25.09	0.08	0.06	40 (87.0%)	0 (0.0%)	18.72	3.04	0.00

#	Model	Backbone	Size	Modality Support	Paradigm
80	NExT-GPT-V1.5 (Wu et al., 2024a)	vicuna-7b-v1.5	7B	Language, Image, Video, Audio	Comprehension+Generation
81	VidAgent† (Shen et al., 2023)	vicuna-7b-v0	7B	Language, Image, Video	Comprehension+Generation

Note that, for VidAgent†, we implement HuggingGPT as the prototype agent, and integrate InternVL-2.5-8B (Chen et al., 2024c) as video comprehension module, and integrate CogVideo (Hong et al., 2022) as video generation module.

C.2 Full Main Evaluation Results

We note that all the generalists run the evaluation on our `General-Bench` data set under a zero-shot setting. The overall results of part of the models on image comprehension and generation are presented in Table 2 and Table 3, respectively; video results are shown in Table 11; audio results are shown in Table 12; 3D results are shown in Table 13; The results of all generalists on NLP tasks are shown in Table 14. The complete performing scores of all MLLMs across all tasks and datasets are presented in Appendix §C. Overall, we have the following observations.

Table 15: Leaderboard of multimodal generalists (MLLMs) at level-2.

Model	Modality	Paradigm	Level 2 Score of Image	of Video	of Audio	of 3D	of Overall	Ranking
Unified-io-2-XXL		C+G	20.62	8.56	25.63	0.00	13.70	1
AnyGPT		C+G	23.10	0.00	29.06	0.00	13.04	2
NExT-GPT-V1.5		C+G	18.69	8.34	25.05	0.00	13.02	3
ImageBind-LLM		C	19.54	12.54	17.52	0.00	12.40	4

Model	Modality	Paradigm	Level 2 Score					Ranking
			of Image	of Video	of Audio	of 3D	of Overall	
ModaVerse-7b-v0		C+G	15.56	7.32	26.10	0.00	12.25	5
Vitron-V1		C+G	30.13	18.72	0.00	0.00	12.21	6
PandaGPT-13B		C	20.78	9.34	16.98	0.00	11.78	7
VidAgent		C+G	18.21	25.00	0.00	0.00	10.80	8
InternVL2_5-8B		C	25.20	8.44	0.00	0.00	8.41	9
Emu2-37B		C+G	30.90	0.00	0.00	0.00	7.73	10
Sa2VA-26B		C	21.88	8.81	0.00	0.00	7.67	11
LaVIT-V2 (7B)		C+G	29.50	0.00	0.00	0.00	7.38	12
LLaVA-One-Vision-72B		C	23.12	5.83	0.00	0.00	7.24	13
Qwen2-Audio-Instruct		C	0.00	0.00	28.61	0.00	7.15	14
Qwen-Audio-Chat		C	0.00	0.00	28.39	0.00	7.10	15
Mini-Gemini		C+G	27.90	0.00	0.00	0.00	6.975	16
SEED-LLaMA-13B		C+G	26.81	0.00	0.00	0.00	6.70	17
GAMA		C	0.00	0.00	26.35	0.00	6.59	18
Qwen2-VL-72B		C	19.41	6.89	0.00	0.00	6.58	19
Sa2VA-8B		C	17.33	8.31	0.00	0.00	6.41	20
InternVL-2.5-26B		C	18.73	6.70	0.00	0.00	6.36	21
Qwen2-VL-7B		C	18.42	6.00	0.00	0.00	6.11	22
InternVL2_5-4B		C	24.41	0.00	0.00	0.00	6.10	23
SALMONN-13B		C	0.00	0.00	23.95	0.00	5.99	24
InternVL-2-26B		C	17.55	6.36	0.00	0.00	5.98	25
WavLLM		C	0.00	0.00	23.49	0.00	5.87	26
Monkey-10B-chat		C	23.51	0.00	0.00	0.00	5.87	27
InternVL2_5-2B		C	23.32	0.00	0.00	0.00	5.83	28
Pengi		C	0.00	0.00	23.29	0.00	5.82	29
LLaVA-One-Vision-7B		C	18.32	4.34	0.00	0.00	5.67	30
SALMONN-7B		C	0.00	0.00	21.09	0.00	5.27	31
InternVL-2.5-8B		C	14.70	5.76	0.00	0.00	5.12	32
DeepSeek-VL-7B-Chat		C	19.89	0.00	0.00	0.00	4.97	33
InternVL-2-8B		C	14.06	5.64	0.00	0.00	4.93	34
GPT4-o		C	19.67	0.00	0.00	0.00	4.92	35
GPT4-o-4096		C	19.68	0.00	0.00	0.00	4.92	36
Gemini-1.5-Pro		C	19.67	0.00	0.00	0.00	4.92	37
Claude-3.5-Sonnet		C	19.38	0.00	0.00	0.00	4.85	38
Claude-3.5-Opus		C	19.00	0.00	0.00	0.00	4.75	39
chatgpt4-o-latest		C	18.98	0.00	0.00	0.00	4.74	40
Gemini-1.5-Flash		C	18.54	0.00	0.00	0.00	4.64	41
CoLVA-4B		C	13.59	4.78	0.00	0.00	4.59	42
GPT4-V		C	18.16	0.00	0.00	0.00	4.54	43
GPT4-o-mini		C	17.79	0.00	0.00	0.00	4.45	44
GLM-VL-Chat		C	17.00	0.00	0.00	0.00	4.25	45
Idefics3-8B-Llama3		C	16.71	0.00	0.00	0.00	4.18	46
LLaVA-NeXT-34B		C	16.58	0.00	0.00	0.00	4.15	47
Phi-3.5-Vision-Instruct		C	16.46	0.00	0.00	0.00	4.12	48
MiniCPM3-4B		C	16.46	0.00	0.00	0.00	4.12	49
CogVLM-Chat		C	16.31	0.00	0.00	0.00	4.08	50
CoLVA-2B		C	11.73	4.47	0.00	0.00	4.05	51
InternVL-Chat-V1-5		C	16.16	0.00	0.00	0.00	4.04	52
DetGPT		C	16.05	0.00	0.00	0.00	4.01	53
BLIP-3 (XGen-MM)		C	15.40	0.00	0.00	0.00	3.85	54
LLaVA-NeXT-13B		C	15.11	0.00	0.00	0.00	3.78	55
Pixtral-12B		C	14.74	0.00	0.00	0.00	3.69	56
ShareGPT4V-13B		C	14.72	0.00	0.00	0.00	3.68	57
Yi-vision-v2		C	14.61	0.00	0.00	0.00	3.65	58
Qwen-VL-Chat		C	13.91	5.34	0.00	0.00	3.48	59
ShareGPT4V-7B		C	13.78	0.00	0.00	0.00	3.45	60

Model	Modality	Paradigm	Level 2 Score					Ranking
			of Image	of Video	of Audio	of 3D	of Overall	
Mini-InternVL-Chat-4B-V1-5		C	13.53	0.00	0.00	0.00	3.38	61
InternLM-XComposer2-VL-1.8B		C	13.31	0.00	0.00	0.00	3.33	62
DeepSeek-VL-7B-Base		C	13.13	0.00	0.00	0.00	3.28	63
MiniGPT4-LLaMA2-7B		C	12.89	0.00	0.00	0.00	3.22	64
MoE-LLAVA-Phi2-2.7B-4e-384		C	12.55	0.00	0.00	0.00	3.14	65
mPLUG-Owl2-LLaMA2-7b		C	12.21	0.00	0.00	0.00	3.05	66
Cambrian-1-8B		C	11.76	0.00	0.00	0.00	2.94	67
BLIP2		C	11.65	0.00	0.00	0.00	2.91	68
miniMonkey		C	11.31	0.00	0.00	0.00	2.83	69
NExT-Chat		C	10.65	0.00	0.00	0.00	2.66	70
Audio-GPT4		G	0.00	0.00	8.80	0.00	2.20	71
GPT4RoI-7B		C	8.49	0.00	0.00	0.00	2.12	72
Show-o		C+G	7.78	0.00	0.00	0.00	1.95	73
SpeechGPT-7B-com		G	0.00	0.00	7.22	0.00	1.81	74
PointLLM-13B		C	0.00	0.00	0.00	7.00	1.75	75
LM4LV		G	0.00	6.74	0.00	0.00	1.69	76
PointLLM-7B		C	0.00	0.00	0.00	6.53	1.63	77
Long-LLaVA-9B		C	10.23	5.84	0.00	0.00	1.46	78
3D-VisTA		C	0.00	0.00	0.00	5.41	1.35	79
3D-LLM-2.1B		C	0.00	0.00	0.00	5.41	1.35	80
OMG-LLaVA-InternLM20B		C	4.56	0.00	0.00	0.00	1.14	81
DeepSeek-VL-2		C	19.21	3.98	0.00	0.00	1.00	82
DeepSeek-VL-2-small		C	17.40	3.64	0.00	0.00	0.91	83
Otter		C	3.15	0.00	0.00	0.00	0.79	84
LLaMA-mesh		G	0.00	0.00	0.00	1.60	0.40	85
LISA		C	1.27	0.00	0.00	0.00	0.32	86
GLaMM		C	0.94	0.00	0.00	0.00	0.24	87
AvatarGPT		C	0.00	0.00	0.00	0.21	0.05	88
MotionGPT-T5		G	0.00	0.00	0.00	0.00	0.00	/
MotionGPT-LLaMA		G	0.00	0.00	0.00	0.00	0.00	/
Meta-Llama-3.1-8B-Instruct		/	0.00	0.00	0.00	0.00	0.00	/
Gemma-2-9b-it		/	0.00	0.00	0.00	0.00	0.00	/
GPT-J		/	0.00	0.00	0.00	0.00	0.00	/
ChatGLM-6B		/	0.00	0.00	0.00	0.00	0.00	/
Qwen2.5-7B-Instruct		/	0.00	0.00	0.00	0.00	0.00	/
InternLM2-Chat-7B		/	0.00	0.00	0.00	0.00	0.00	/
Baichuan2-7B-Base		/	0.00	0.00	0.00	0.00	0.00	/
Vicuna-7b-V1.5		/	0.00	0.00	0.00	0.00	0.00	/
Falcon3-7B-Instruct		/	0.00	0.00	0.00	0.00	0.00	/
Ministral-8B-Instruct-2410		/	0.00	0.00	0.00	0.00	0.00	/
Yi-lightning		/	0.00	0.00	0.00	0.00	0.00	/
GPT-3.5-turbo		/	0.00	0.00	0.00	0.00	0.00	/

Table 17: Leaderboard of multimodal generalists (MLLMs) at level-3 where Comprehension and Generation.

Model	Modality	Paradigm	Level 3 Score					Ranking
			of Image	of Video	of Audio	of 3D	of Overall	
Sa2VA-26B		C	14.65	4.58	0.00	0.00	4.81	1
LLaVA-One-Vision-72B		C	15.21	3.75	0.00	0.00	4.74	2
Qwen2-VL-72B		C	12.34	5.22	0.00	0.00	4.39	3
Mini-Gemini		C+G	17.23	0.00	0.00	0.00	4.31	4
Sa2VA-8B		C	12.39	4.38	0.00	0.00	4.19	5
InternVL2_5-8B		C	13.09	1.82	0.00	0.00	3.73	6
GPT4-o-4096		C	14.68	0.00	0.00	0.00	3.67	7
Qwen2-VL-7B		C	12.13	2.47	0.00	0.00	3.65	8

Model	Modality	Paradigm	Level 3 Score					Ranking
			of Image	of Video	of Audio	of 3D	of Overall	
GPT4-o		C	14.51	0.00	0.00	0.00	3.63	9
InternVL-2-26B		C	8.81	4.81	0.00	0.00	3.41	10
InternVL-2.5-26B		C	9.51	3.76	0.00	0.00	3.32	11
ChatGPT-o-latest		C	13.02	0.00	0.00	0.00	3.26	12
GPT4-V		C	12.85	0.00	0.00	0.00	3.21	13
Gemini-1.5-Pro		C	12.66	0.00	0.00	0.00	3.17	14
Claude-3.5-Sonnet		C	11.98	0.00	0.00	0.00	3.00	15
GPT4-o-mini		C	11.94	0.00	0.00	0.00	2.99	16
LLaVA-One-Vision-7B		C	10.21	1.54	0.00	0.00	2.94	17
InternVL2_5-4B		C	11.59	0.00	0.00	0.00	2.90	18
Monkey-10B-chat		C	11.59	0.00	0.00	0.00	2.90	19
InternVL2_5-2B		C	11.45	0.00	0.00	0.00	2.86	20
Claude-3.5-Opus		C	11.08	0.00	0.00	0.00	2.77	21
Gemini-1.5-Flash		C	10.85	0.00	0.00	0.00	2.71	22
Vitron-V1		C+G	7.65	3.04	0.00	0.00	2.67	23
CoLVA-4B		C	9.45	1.24	0.00	0.00	2.67	24
Qwen-Audio-Chat		C	0.00	0.00	10.57	0.00	2.64	25
InternVL-Chat-V1-5		C	9.42	0.00	0.00	0.00	2.36	26
Phi-3.5-Vision-Instruct		C	9.39	0.00	0.00	0.00	2.35	27
DeepSeek-VL-2		C	8.32	0.64	0.00	0.00	2.24	28
InternVL-2.5-8B		C	7.63	1.24	0.00	0.00	2.22	29
GLM-VL-Chat		C	8.67	0.00	0.00	0.00	2.17	30
Qwen2-Audio-Instruct		C	0.00	0.00	8.53	0.00	2.13	31
LLaVA-NeXT-34B		C	8.24	0.00	0.00	0.00	2.06	32
DeepSeek-VL-7B-Chat		C	8.19	0.00	0.00	0.00	2.05	33
MiniCPM3-4B		C	8.11	0.00	0.00	0.00	2.03	34
Long-LLaVA-9B		C	4.21	3.81	0.00	0.00	2.01	35
Yi-vision-v2		C	7.85	0.00	0.00	0.00	1.96	36
CogVLM-Chat		C	7.77	0.00	0.00	0.00	1.94	37
InternVL-2-8B		C	7.28	0.46	0.00	0.00	1.94	38
Idefics3-8B-Llama3		C	7.70	0.00	0.00	0.00	1.93	39
CoLVA-2B		C	6.60	1.04	0.00	0.00	1.91	40
GAMA		C	0.00	0.00	7.15	0.00	1.79	41
LLaVA-NeXT-13B		C	6.87	0.00	0.00	0.00	1.72	42
BLIP-3 (XGen-MM)		C	6.42	0.00	0.00	0.00	1.61	43
ShareGPT4V-13B		C	5.97	0.00	0.00	0.00	1.49	44
Qwen-VL-Chat		C	5.88	0.00	0.00	0.00	1.47	45
DeepSeek-VL-7B-Base		C	5.75	0.00	0.00	0.00	1.44	46
Pixtral-12B		C	5.72	0.00	0.00	0.00	1.43	47
DeepSeek-VL-2-small		C	5.12	0.52	0.00	0.00	1.41	48
MoE-LLAVA-Phi2-2.7B-4e-384		C	5.47	0.00	0.00	0.00	1.37	49
NExT-GPT-V1.5		C+G	3.24	0.71	1.34	0.00	1.32	50
Mini-InternVL-Chat-4B-V1-5		C	5.21	0.00	0.00	0.00	1.30	51
Emu2-37B		C+G	5.18	0.00	0.00	0.00	1.30	52
InternLM-XComposer2-VL-1.8B		C	4.78	0.00	0.00	0.00	1.20	53
ShareGPT4V-7B		C	4.78	0.00	0.00	0.00	1.20	54
MiniGPT4-LLaMA2-7B		C	4.68	0.00	0.00	0.00	1.17	55
mPLUG-Owl2-LLaMA2-7b		C	4.60	0.00	0.00	0.00	1.15	56
AnyGPT		C+G	1.29	0.00	3.29	0.00	1.15	57
miniMonkey		C	4.51	0.00	0.00	0.00	1.13	58
Cambrian-1-8B		C	3.84	0.00	0.00	0.00	0.96	59
DetGPT		C	3.77	0.00	0.00	0.00	0.94	60
LaVIT-V2 (7B)		C+G	3.71	0.00	0.00	0.00	0.93	61
SALMONN-13B		C	0.00	0.00	3.61	0.00	0.90	62
ImageBind-LLM		C	1.56	0.72	1.26	0.00	0.89	63
NExT-Chat		C	3.51	0.00	0.00	0.00	0.88	64

Model	Modality	Paradigm	Level 3 Score					Ranking
			of Image	of Video	of Audio	of 3D	of Overall	
SEED-LLaMA-13B		C+G	3.49	0.00	0.00	0.00	0.87	65
WavLLM		C	0.00	0.00	3.28	0.00	0.82	66
Unified-io-2-XXL		C+G	2.11	0.14	1.01	0.00	0.82	67
ModaVerse-7b-v0		C+G	0.98	0.23	1.14	0.78	0.78	68
PandaGPT-13B		C	2.35	0.05	0.65	0.00	0.76	69
Audio-GPT4		G	0.00	0.00	3.02	0.00	0.76	70
BLIP2		C	2.79	0.00	0.00	0.00	0.70	71
GPT4RoI-7B		C	2.36	0.00	0.00	0.00	0.59	72
Pengi		C	0.00	0.00	1.74	0.00	0.44	73
3D-LLM-2.1B		C	0.00	0.00	0.00	1.38	0.35	74
3D-VisTA		C	0.00	0.00	0.00	1.07	0.27	75
Show-o		C+G	0.84	0.00	0.00	0.00	0.21	76
LISA		C	0.82	0.00	0.00	0.00	0.21	77
Otter		C	0.68	0.00	0.00	0.00	0.17	78
OMG-LLaVA-InternLM20B		C	0.44	0.00	0.00	0.00	0.11	79
GLaMM		C	0.41	0.00	0.00	0.00	0.10	80
AvatarGPT		C	0.00	0.00	0.00	0.21	0.05	81
PointLLM-7B		C	0.00	0.00	0.00	0.00	0.00	/
PointLLM-13B		C	0.00	0.00	0.00	0.00	0.00	/
MotionGPT-T5		G	0.00	0.00	0.00	0.00	0.00	/
MotionGPT-LLaMA		G	0.00	0.00	0.00	0.00	0.00	/
LLaMA-mesh		G	0.00	0.00	0.00	0.00	0.00	/
SALMONN-7B		C	0.00	0.00	0.00	0.00	0.00	/
SpeechGPT-7B-com		G	0.00	0.00	0.00	0.00	0.00	/
LM4LV		G	0.00	0.00	0.00	0.00	0.00	/
VidAgent		C+G	0.00	0.00	0.00	0.00	0.00	/
Meta-Llama-3.1-8B-Instruct		/	0.00	0.00	0.00	0.00	0.00	/
Gemma-2-9b-it		/	0.00	0.00	0.00	0.00	0.00	/
GPT-J		/	0.00	0.00	0.00	0.00	0.00	/
ChatGLM-6B		/	0.00	0.00	0.00	0.00	0.00	/
Qwen2.5-7B-Instruct		/	0.00	0.00	0.00	0.00	0.00	/
InternLM2-Chat-7B		/	0.00	0.00	0.00	0.00	0.00	/
Baichuan2-7B-Base		/	0.00	0.00	0.00	0.00	0.00	/
Vicuna-7b-V1.5		/	0.00	0.00	0.00	0.00	0.00	/
Falcon3-7B-Instruct		/	0.00	0.00	0.00	0.00	0.00	/
Ministral-8B-Instruct-2410		/	0.00	0.00	0.00	0.00	0.00	/
Yi-lightning		/	0.00	0.00	0.00	0.00	0.00	/
GPT-3.5-turbo		/	0.00	0.00	0.00	0.00	0.00	/

C.3 Level and Leaderboard of Multimodal Generalists

Based on the overall performance of each model across the various modalities and tasks, we rank all the compared models according to the `General-Level` scoring defined in § 3.2. Tables 15, 17 and 19 present the specific scores and rankings of multimodal generalists at different General-Levels. Note that no generalists score non-zero at Level-5, and thus we do not show a rank at Level-5. Figure 4 visualizes these leaderboards.

As shown, for all the current MLLMs at level 2, Unified-io-2-XXL (Lu et al., 2024a) ranks the best, followed by AnyGPT (Zhan et al., 2024). Surprisingly, GPT-4V and GPT-4o did not achieve the expected rankings at level 2. While the GPT series excels in the individual tasks it supports, as generalists, they fall short in skill coverage compared to some open-source MLLMs. This is because, to rank higher at level 2, models must not only perform well on different tasks but also support as many modalities and tasks as possible.

Next, MLLMs that can manage to reach level 3 become different. Sa2VA-26B (Yuan et al., 2025) ranks at the top, while LLaVA-One-Vision-72B (Li et al., 2024d) and Qwen2-VL-72B (Wang et al., 2024a) achieve the second and third places, respectively. Some high-ranking level-2 models lost their places at level 3. This lies in the fact that most MLLMs are limited

Table 12: Performance of multimodal generalists on audio comprehension and generation skills.

Model	Audio Comprehension Skill (Avg within each #A-C Group)									Task Completion		Level Score on Audio		
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	87.27	79.08	70.62	79.00	71.87	62.90	58.70	77.90	78.07	/	/	/	/	/
Qwen-Audio-Chat	56.93	68.77	76.80	37.70	47.71	19.79	56.44	85.15	78.50	30 (100.0%)	6 (25.0%)	28.39	10.57	0.00
Qwen2-Audio-Instru	72.65	74.80	61.40	36.80	45.82	13.45	61.68	78.95	67.99	24 (100.0%)	6 (25.0%)	28.61	8.53	0.00
GAMA	57.00	64.20	68.00	53.20	18.43	26.95	48.85	85.55	61.80	23 (95.8%)	4 (16.7%)	26.35	7.15	0.00
Pengi	52.88	60.07	56.70	36.78	19.77	19.55	42.95	77.40	61.17	23 (95.8%)	1 (4.2%)	23.29	1.74	0.00
SALMONN-13B	67.89	56.33	67.80	29.45	24.67	19.36	43.95	76.55	56.67	23 (95.8%)	2 (8.3%)	23.95	3.61	0.00
WavLLM	64.45	41.07	71.20	30.08	31.30	26.55	45.75	61.40	64.57	24 (100.0%)	2 (8.3%)	23.49	3.28	0.00
NExT-GPT-V1.5	43.23	29.13	65.80	26.70	14.47	25.65	47.95	70.20	69.43	24 (100.0%)	0 (0.0%)	25.05	1.34	0.00
PandaGPT (13B)	41.80	20.23	45.20	20.98	8.47	20.50	42.25	54.80	65.83	24 (100.0%)	0 (0.0%)	16.98	0.65	0.00
ModaVerse-7b-v0	34.10	16.37	32.80	15.20	6.60	8.90	35.05	49.20	60.13	23 (95.8%)	0 (0.0%)	26.10	1.14	0.00
Any-GPT	44.50	32.13	63.40	48.08	16.27	36.40	52.65	67.95	44.63	23 (95.8%)	1 (4.2%)	29.06	3.29	0.00
Unified-io-2-XXL	30.15	27.60	56.10	28.58	15.47	38.35	38.70	63.50	60.63	24 (100.0%)	0 (0.0%)	25.63	1.01	0.00

Model	Audio Generation Skill (Avg within each #A-G Group)											Task Completion		Level Score on Audio		
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	31.50	3.82	3.64	4.68	41.54	51.40	11.52	6.80	8.33	22.88	20.33	/	/	/	/	/
Unified-io-2-XXL	18.36	2.03	5.11	40.52	16.41	24.31	16.97	86.23	94.52	0.25	2.24	17 (85.0%)	0 (0.0%)	25.63	1.01	0.00
Any-GPT	23.50	3.24	4.57	33.58	13.38	14.05	27.49	45.36	83.89	0.25	2.47	17 (85.0%)	1 (5.0%)	29.06	3.29	0.00
NExT-GPT-V1.5	13.60	1.15	4.07	50.51	34.51	1.35	12.36	96.70	99.23	0.25	7.77	17 (85.0%)	1 (5.0%)	25.05	1.34	0.00
AudioGPT	0.50	1.32	4.61	23.10	29.48	0.00	0.00	46.30	79.98	0.25	0.00	13 (65.0%)	1 (5.0%)	8.80	3.02	0.00
SpeechGPT	0.10	2.79	4.44	32.35	0.00	0.00	0.00	30.24	85.54	0.25	0.00	11 (55.0%)	0 (0.0%)	7.22	0.00	0.00
ModaVerse	12.30	1.15	4.29	50.50	28.99	1.05	16.45	100.00	100.00	0.25	4.17	17 (85.0%)	2 (10.0%)	26.10	1.14	0.00

Table 13: Performance of multimodal generalists on 3D comprehension and generation skills.

Model	3D Comprehension Skill (Avg within each #D-C Group)													Task Completion		Level Score on 3D		
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	96.24	98.35	97.78	78.50	70.02	81.20	55.00	88.28	75.20	9.96	68.52	47.14	22.30	/	/	/	/	/
3D-VisTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	46.37	0.00	7 (23.3%)	2 (6.7%)	5.41	1.07	0.00
PointLLM-7B	46.16	7.50	72.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8 (26.7%)	0 (0.0%)	6.53	0.00	0.00
PointLLM-13B	48.79	10.00	78.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9 (30.0%)	0 (0.0%)	7.00	0.00	0.00
3D-LLM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	46.34	0.00	7 (23.3%)	1 (3.3%)	5.41	1.38	0.00
AvatarGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.70	1 (3.3%)	0 (0.0%)	0.21	0.21	0.00

Model	3D Generation Skill (Avg within each #D-G Group)									Task Completion		Level Score on 3D		
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	0.22	7.12E-5	24.42	25.69	78.06	83.64	6540.02	6540.02	0.23	/	/	/	/	/
MotionGPT-T5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.51	1 (4.5%)	0 (0.0%)	0.00	0.00	0.00
MotionGPT-LLaMA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60	1 (4.5%)	0 (0.0%)	0.00	0.00	0.00
LLaMA-Mesh	0.00	0.00	0.00	17.55	0.00	0.00	0.00	0.00	0.00	1 (4.5%)	0 (0.0%)	1.60	0.00	0.00

to multimodal content comprehension and lack support for generation tasks. GPT-4V and GPT-4o win top-10 positions here.

Finally, only 3 MLLMs that reach level 4 can be seen, i.e., Mini-Gemini, Emu2-37B, and Vitron-V1. At this level, these models exhibit synergy across both comprehension and generation. Besides these three models, no other systems exhibit such capability.

Most critically, no model has yet demonstrated the ability to enhance language intelligence through non-language modalities, underscoring the significant challenges in the pursuit of true AGI. And this is definitely our goal to reach the most capable multimodal generalists.

C.4 Capability BreakDown

We now take a closer look, as multimodal generalists, at how well different MLLMs support tasks and modalities.

Table 14: Performance of generalists on language-related (NLP) skills.

Model	Language Skill (Avg within each #L Group) #1 #12	#2 #13	#3 #14	#4 #15	#5 #16	#6 #17	#7 #18	#8 #19	#9 #20	#10 #21	#11 #22	Task Completion #Supported Task	Task Completion #Win-over-Specialist	Level Score Level-5
SoTA Specialist	62.62 86.95	86.23 0.31	76.78 94.40	71.00 91.41	58.02 86.05	62.80 86.03	75.11 84.72	77.84 83.67	79.70 58.61	71.91 77.73	28.27 92.38	/	/	/
Meta-Llama-3.1-8B-Instruct	39.75 45.34	56.76 7.95	54.21 76.40	60.52 51.80	20.01 65.90	37.17 41.10	36.23 24.49	29.12 30.70	53.23 8.08	44.49 32.40	14.80 54.35	113 (98.3%)	0 (0.0%)	0.00
ChatGLM-6b	28.97 42.84	33.24 10.91	37.24 41.80	46.10 45.81	19.39 24.50	27.84 16.45	18.85 0.12	35.88 8.41	27.85 2.70	38.51 23.80	13.93 45.37	96 (83.5%)	0 (0.0%)	0.00
Vicuna-7b-v1.5	24.78 43.98	11.18 11.41	33.44 0.00	41.19 0.00	4.51 0.00	13.25 0.96	19.94 0.07	35.27 0.47	54.81 0.00	40.58 23.13	5.06 15.40	72 (62.6%)	0 (0.0%)	0.00
Falcon3-7B-Instruct	36.79 48.15	58.36 5.15	49.91 88.80	56.80 85.89	21.38 45.65	37.12 42.86	32.03 27.64	42.11 34.22	55.79 11.19	42.07 39.80	15.56 58.75	112 (97.4%)	0 (0.0%)	0.00
Ministral-8B-Instruct-2410	41.74 23.39	54.21 11.08	49.53 84.80	51.92 72.60	39.32 56.70	40.49 37.14	13.00 6.28	22.86 31.38	56.87 9.37	43.46 25.53	13.73 40.44	112 (97.4%)	0 (0.0%)	0.00
Yi-Lightning	41.73 52.68	60.54 5.37	55.39 72.60	60.51 56.24	20.53 64.75	39.83 43.59	22.45 28.27	43.57 42.84	62.52 25.34	42.03 29.27	15.29 60.49	113 (98.3%)	0 (0.0%)	0.00
GPT-4V	27.55 44.56	62.40 3.16	34.57 86.20	32.55 83.23	14.43 65.10	27.84 53.82	27.79 54.14	36.07 45.45	65.36 33.86	42.11 26.46	13.96 24.24	113 (98.3%)	0 (0.0%)	0.00
GPT-4o	26.25 46.41	62.57 2.58	33.98 85.40	31.50 86.30	16.20 67.50	26.26 56.10	27.14 57.42	36.64 46.97	66.86 39.52	42.69 32.07	14.49 28.50	113 (98.3%)	0 (0.0%)	0.00
Emu2-32B	32.91 50.15	45.43 9.53	47.04 57.54	39.56 48.78	27.74 43.76	31.24 36.67	39.04 19.84	41.72 24.01	45.48 13.78	46.35 26.47	13.05 31.72	113 (98.3%)	0 (0.0%)	0.00
DeepSeek-VL-7B	29.97 79.68	44.39 83.00	55.55 62.20	20.36 50.60	40.49 62.30	57.93 46.87	49.85 4.12	48.73 28.46	27.03 8.11	56.76 31.80	10.37 40.97	114 (99.1%)	0 (0.0%)	0.00
Qwen2-VL-7B	23.91 37.23	27.51 6.48	37.68 64.00	46.40 37.00	17.84 3.50	20.96 20.50	36.25 0.24	29.29 4.87	35.42 6.00	35.58 20.87	12.62 21.79	94 (81.7%)	0 (0.0%)	0.00
LLaVA-One-Vision-72B	50.44 43.81	41.98 3.55	54.55 84.80	61.13 10.43	29.87 59.35	56.99 34.91	35.24 42.94	43.27 28.63	55.23 19.26	41.49 52.20	17.73 71.95	110 (95.7%)	0 (0.0%)	0.00
InternVL2.5-8B	42.93 71.96	47.76 75.20	59.54 55.40	31.17 68.40	42.86 56.75	32.72 55.60	50.98 22.12	43.02 36.48	30.85 9.80	51.23 32.13	9.07 53.67	114 (99.1%)	0 (0.0%)	0.00
Long-llava	26.50 48.44	49.49 11.40	34.81 68.60	39.62 41.70	17.83 52.65	33.14 31.42	20.63 2.33	38.44 21.52	48.90 7.40	38.10 29.47	6.30 42.07	107 (93.0%)	0 (0.0%)	0.00
NExT-GPT-V1.5	20.66 42.09	22.42 1.06	32.55 68.90	39.51 43.20	4.19 28.78	16.47 9.24	16.49 4.44	32.67 6.16	51.49 7.22	37.73 24.17	5.06 18.86	79 (68.7%)	0 (0.0%)	0.00
SEED-LLaMA-13B	18.11 20.84	32.55 11.16	26.54 13.20	25.19 34.80	8.80 28.98	18.85 19.93	11.68 2.59	15.89 10.31	21.64 2.10	23.56 12.07	4.80 12.41	109 (94.8%)	0 (0.0%)	0.00
LLaMA-Mesh	29.34 44.19	16.70 11.41	47.05 0.00	56.85 0.00	5.09 0.00	14.85 1.50	21.27 0.65	39.24 0.56	57.40 0.01	40.65 23.13	4.93 21.57	84 (73.0%)	0 (0.0%)	0.00
MiniGPT4-LLaMA2	28.17 42.56	15.73 7.46	40.98 0.00	45.15 0.00	3.71 0.00	10.99 2.08	25.97 6.36	43.03 4.52	35.22 0.00	36.14 17.55	10.42 21.23	84 (73.0%)	0 (0.0%)	0.00

Table 19: Leaderboard of multimodal generalists (MLLMs) at level-4, where Comprehension and Generation.

Model	Modality	Paradigm	Level 4 Score of Image	of Video	of Audio	of 3D	of Overall	Ranking
Mini-Gemini		C+G	6.23	0.00	0.00	0.00	1.56	1
Vitron-V1		C+G	4.59	0.00	0.00	0.00	1.15	2
Emu2-37B		C+G	1.25	0.00	0.00	0.00	0.31	3

Task Supporting. In Figure 25, we present all skills (meta-tasks) supported by different MLLMs across various modalities and within the scopes of comprehension and generation. Overall, MLLMs show relatively lower task support for 3D tasks and skills, compared with the status for other modalities. Also, the coverage of comprehension-related skills by MLLMs should be generally higher than that of generation-related skills. This trend is consistent with the results observed in previous experiments. A significant trend we identified is that MLLMs tend to support skills within only one (or a few) task paradigms. This results in differentiated skill support across different MLLMs, with few models capable of supporting a wide range of skills across diverse tasks.

Moreover, most MLLMs are inclined to focus on basic skills or tasks with simpler and more straightforward definitions,

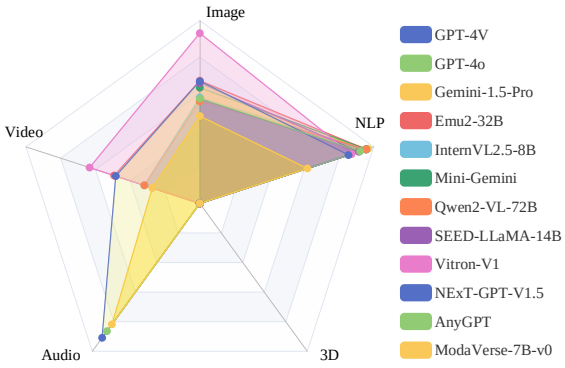


Figure 20: Supporting modality.

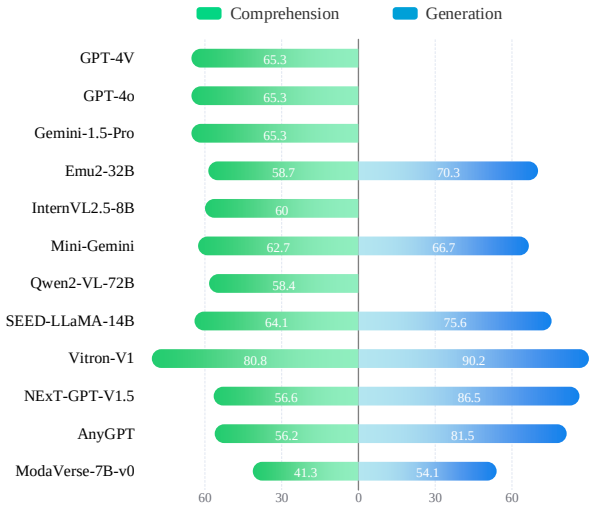


Figure 21: Supporting generation and comprehension.

while tasks requiring complex content output and advanced skills are supported by far fewer models. For instance, compared to coarse-grained visual understanding tasks (e.g., captioning and classification), tasks with more complex definitions—such as pixel-level object detection, image/3D segmentation, video tracking, and image generation—are supported by far fewer existing MLLMs. However, we observe that a few MLLMs stand out for their broader support of cross-modal skills, such as Vitron-V1 (Fei et al., 2024a). Thanks to their architectural designs, these models demonstrate a wider range of task support compared to others. We emphasize that supporting as many task paradigms as possible is a critical requirement for developing more capable multimodal generalists.

Modality Supporting. The broader the range of supported modalities, the more general and versatile the model's capabilities are. Our benchmark emphasizes the evaluation of MLLMs' all-modality capabilities. As shown in the experimental results above, most MLLMs support only a single modality (excluding the language modality, which is inherently supported by LLMs). To further illustrate this, Figure 20 compares the multimodal support capabilities of several top-performing MLLMs. In general, there are very few MLLMs capable of supporting multiple modalities simultaneously. In most cases, MLLMs support one non-language modality, e.g., GPT-4V (OpenAI, 2022b), Emu2-32B (Sun et al., 2024), Mini-Gemini (Li et al., 2024c), InternVL2.5-8B (Chen et al., 2024c). Only a few MLLMs stand out with broader cross-modal or even all-modality support capabilities, encompassing language, image, video, and audio modalities. Examples include NExT-GPT-V1.5 (Wu et al., 2024a), Unified-io-2-XXL (Lu et al., 2024a), and AnyGPT (Zhan et al., 2024), etc. Thanks to their architectural designs, these systems demonstrate a wider range of modality and task support compared to others.

Capabilities on Comprehension vs. Generation. Within a single modality, tasks and skills can be categorized into comprehension and generation types. Here, we explore the capabilities of different MLLMs in supporting these two paradigms. We select several representative MLLMs that can or cannot support both comprehension and generation within various modalities for comparison. Figure 21 directly presents the statistics on these models' capabilities in these two aspects. Overall, support for content comprehension significantly outweighs support for content generation. This phenomenon aligns with practical realities, as modeling tasks for comprehension typically involve expressing the understood content in language, which is relatively straightforward. In contrast, generating multimodal content requires additional efforts in model decoding and extra training, making it a more challenging capability to achieve. It is also evident that different models exhibit varying balances between comprehension and generation capabilities. Among them, Vitron-V1 (Fei et al., 2024a) demonstrates the most comprehensive and well-rounded capabilities in both comprehension and generation to date, i.e., supporting the largest number of both paradigms.

C.5 Analysis and Discussion on Synergy

Next, we conduct a finer-grained analysis of the synergy performance of various MLLMs.

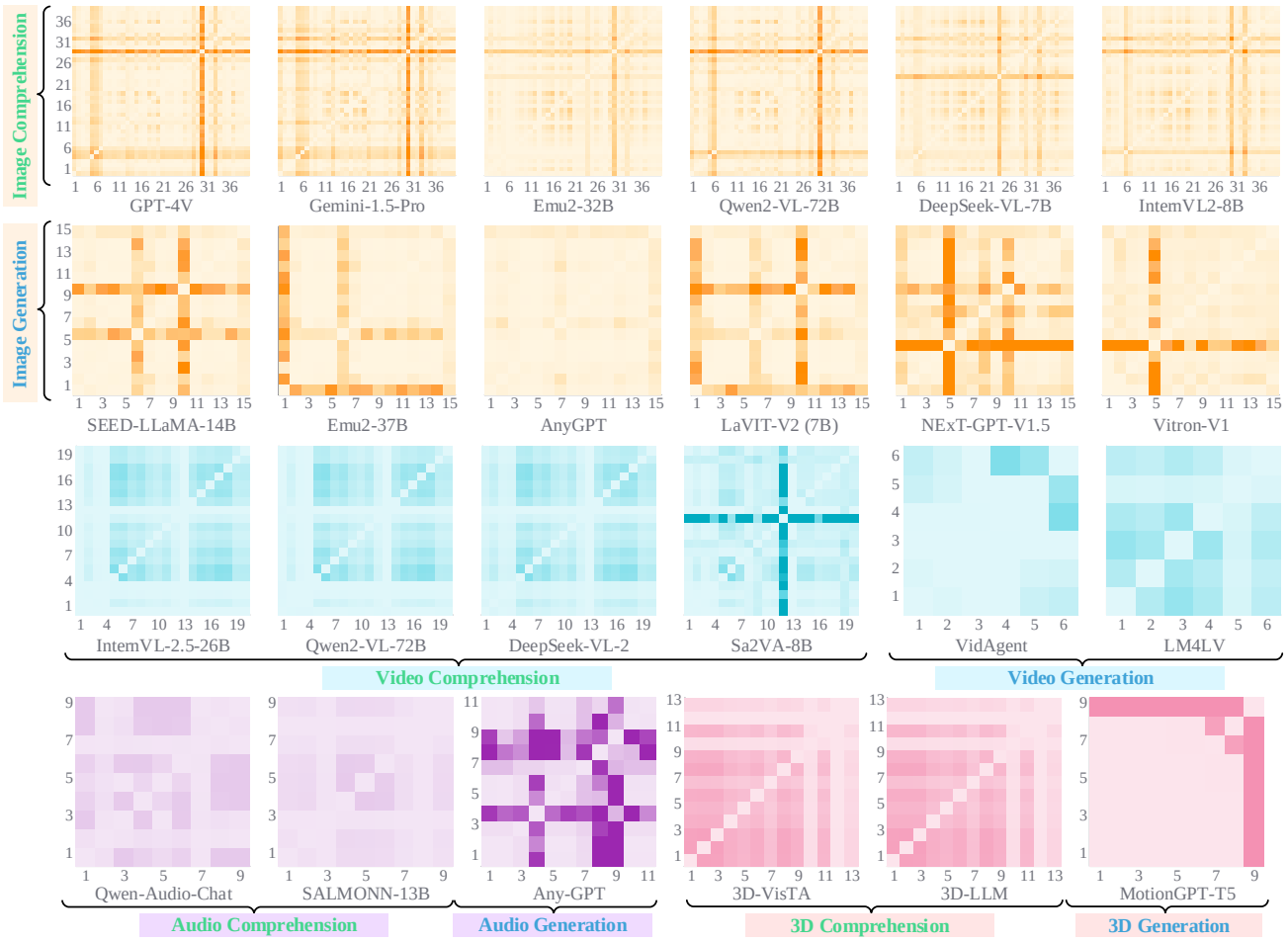


Figure 22: Visualizations of synergy effects between all different skills of various MLLMs.

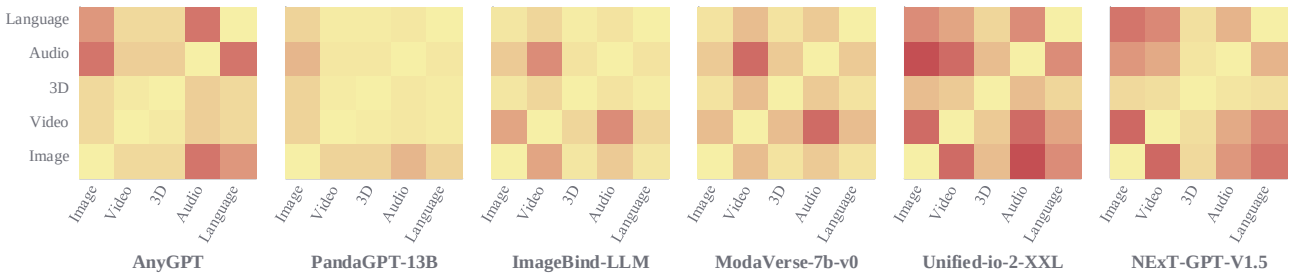


Figure 23: Visualizations (symmetrised) of synergy effects between modalities of various MLLMs.

Synergy Across Skills. First, we examine the different synergy effects exhibited by models across various skills. Skills are categorized based on different modalities and further divided into comprehension and generation categories. We explore the synergy effects displayed by top-performing MLLMs within these skill groups. Technically, we calculate the synergy score for tasks (skills) based on the level-3 score algorithm from `General-Level` (§ 3.2). Then we count the number of synergy tasks for each model and use scores exceeding SoTA specialists as weights. Figure 22 visualizes the results with heatmaps. It is shown that different models exhibit varying levels of cross-task synergy capabilities. Overall, models that achieve higher level-3 scores tend to display denser clusters of highlighted cells in the heatmap. Also, we observe that synergy effects are more likely to occur among closely related skills, as knowledge and information are more easily transferable between similar tasks. This trend is consistent across different modalities. Furthermore, generation tasks seem to exhibit stronger synergy effects compared to tasks within the comprehension category.

Synergy Across Modalities. Finally, we analyze whether different models have learned synergy effects across modalities. The approach is similar to the previous methods, where we calculate instances where performance across modalities exceeds that of SoTA specialists. Figure 23 visualizes the performance of 6 representative strong-performing MLLMs that cover as many modalities as possible. Several notable trends emerge. First, we observe that there is significant synergy occurs between the image and video modalities. This is reasonable, as static image information and dynamic video content are both fundamentally visual in nature, allowing for substantial information sharing that enhances performance across tasks.

Most strikingly, although language appears to exhibit a synergy effect with various other modalities, this effect is in fact unidirectional, specifically from language $\rightarrow$ other modalities. Based on our previous results on NLP tasks, no synergy has been observed from any other modality to the language modality. While in theory, the audio modality should be closely related to language, and we would expect to observe synergy between audio and language tasks, this is not reflected in current results. This limitation is likely due to the reliance of audio-based LLM architectures on the language intelligence of LLMs, which has not yet translated into performance improvements that exceed those of NLP SoTA specialists. We thus strongly urge the MLLM community to prioritize enhancing cross-modality synergy capabilities to advance the development of truly comprehensive multimodal generalists.

Synergy Across Comprehension and Generation. We further investigate the synergy across comprehension and generation, which represents a broader type of synergy than at the task level. Using a similar approach, we calculate the synergy score based on the level-4 score algorithm from `General-Level`. Specifically, we count the frequency of synergy occurrences for each model and use the performance increments exceeding SoTA specialists as weights. The total sum (after normalization) constitutes the synergy score across comprehension and generation. Figure 24 presents the comparisons for only 3 models. As shown, Mini-Gemini (Li et al., 2024c) demonstrates the best synergy effects at this level, securing the top rank on our `General-Level` leaderboard. Also, we observe that the cross-comprehension-generation synergy is only pronounced in the image modality, as these 3 models all support only image modalities.

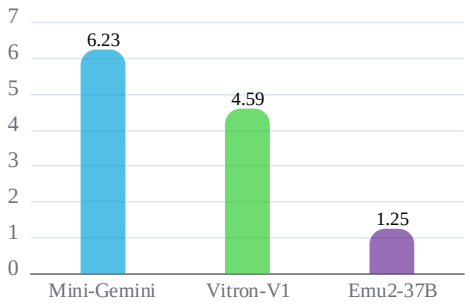


Figure 24: Synergy strength between comprehension and generation.

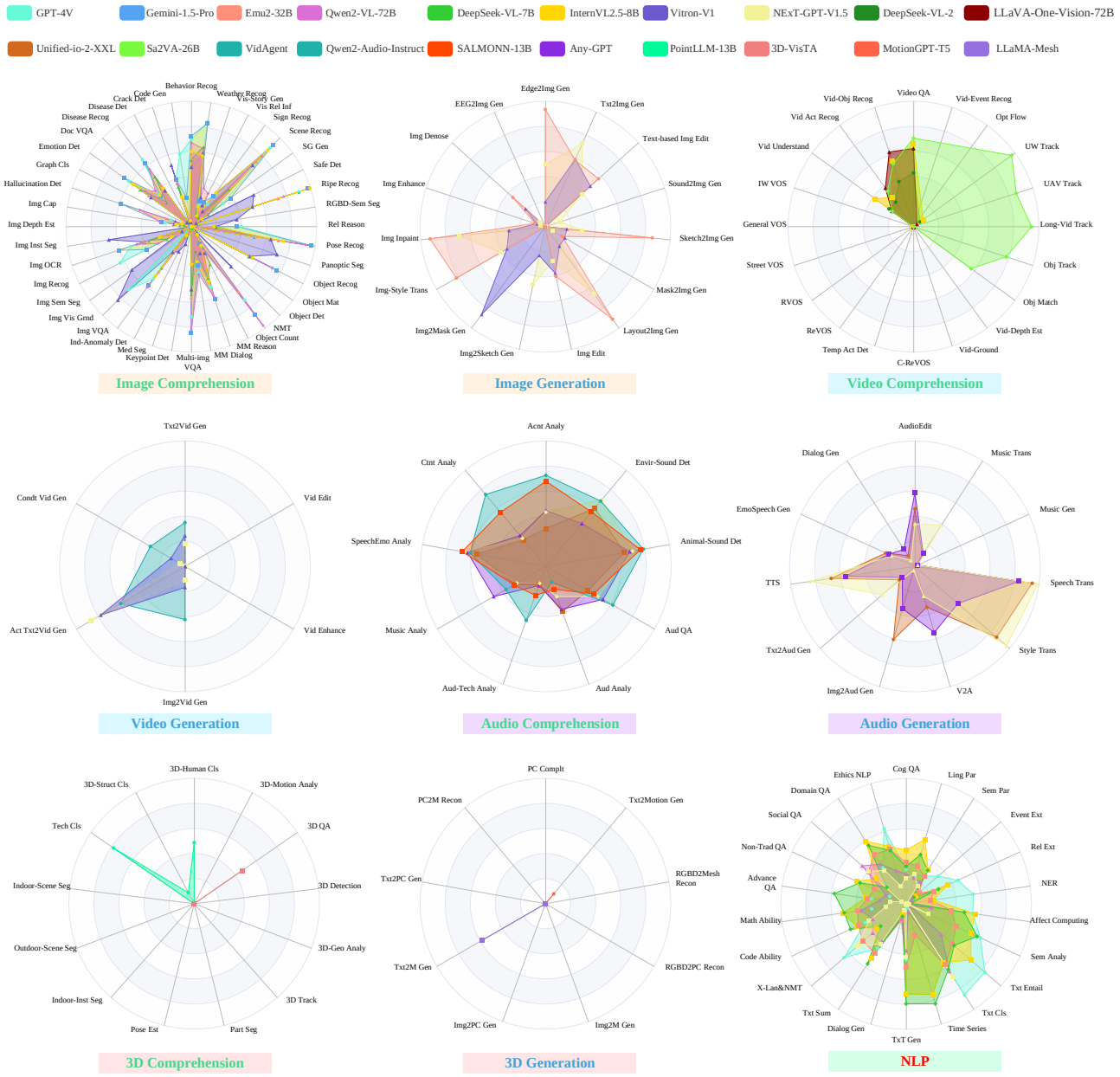


Figure 25: Visualization of skill support in various multimodal generalists.

C.6 Results of Image-related Tasks

Image Comprehension Results. The complete results of all models on image comprehension are presented in Table 20 to Table 38. For the full list of results of all tasks, please visit the arXiv full version of the paper: https://arxiv.org/abs/2505.04620.

Table 20: Results on Image Comprehension Group, from #I-C-1 to #I-C-5. Scores over SoTA specialists are bolded.

Model	#I-C-1 (Behavior Recog)			#I-C-2 (Code Gen)	#I-C-3 (Crack Det)		#I-C-4 (Disease Det)	#I-C-5 (Disease Recog)					
	#1 ↑	#2 ↑	#3 ↑	#1 ↑	#1 ↑	#2 ↑	#1 ↑	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑
SoTA Specialist	21.70	82.30	49.80	53.32	63.08	21.00	22.30	40.52	40.50	35.90	38.42	16.40	62.40
GPT4V	**57.00**	**87.97**	**63.30**	**58.64**	**79.08**	0.00	0.00	**56.60**	**99.75**	**90.92**	38.01	**49.40**	62.40
GPT4o	**58.70**	**90.10**	**72.80**	**63.42**	**86.46**	0.00	0.00	**69.25**	**98.64**	**94.97**	**50.89**	**53.00**	**62.60**
GPT4o-mini	**25.70**	**88.50**	**70.20**	**61.38**	**77.85**	0.00	0.00	**75.00**	**96.93**	**50.14**	**44.55**	**40.40**	49.60
GPT-4o-4096	**24.90**	**90.00**	**89.40**	0.00	**89.54**	0.00	0.00	**74.43**	**99.75**	**99.88**	**63.56**	**52.00**	62.40
ChatGPT-4o-latest	**29.60**	66.90	**80.40**	**57.28**	**85.54**	0.00	0.00	**63.22**	**98.40**	**97.23**	37.21	**47.20**	59.34
Claude-3.5-Sonnet	**46.67**	**87.90**	**68.60**	**60.86**	**80.30**	0.00	0.00	**66.32**	**97.72**	**78.40**	**44.37**	**47.38**	57.23
Claude-3.5-Opus	**43.16**	**88.58**	**64.40**	**57.69**	**79.90**	0.00	0.00	**63.10**	**96.64**	**76.74**	**43.48**	**45.40**	54.76
Emu2-32B	**32.11**	82.26	46.91	7.31	**73.23**	0.00	0.00	**41.67**	40.50	35.90	**40.99**	**38.40**	50.40
DetGPT	**24.49**	76.13	**55.92**	0.00	59.07	16.08	3.25	31.89	39.98	28.73	31.09	**26.20**	44.60
InternVL2.5-8B	**27.10**	73.57	**79.20**	4.86	49.85	0.00	0.00	**48.56**	**53.44**	6.15	30.11	**27.20**	**63.00**
InternVL2.5-4B	**32.30**	63.77	39.80	4.97	34.15	0.00	0.00	**60.06**	**46.31**	21.65	36.63	**25.20**	62.20
InternVL2.5-2B	**26.80**	76.73	**61.80**	4.64	**66.22**	0.00	0.00	37.36	**90.79**	23.04	**39.41**	**31.20**	60.40
Monkey-10B-chat	**24.60**	**83.03**	47.80	0.00	37.23	0.00	0.00	36.49	**94.59**	0.00	28.12	**38.00**	62.40
DeepSeek-VL-7B-Chat	20.00	79.70	**60.20**	4.27	**67.57**	0.00	0.00	37.64	**100.00**	19.57	**43.56**	**28.20**	62.40
Qwen2-VL-7B	20.00	**85.13**	**89.40**	3.61	58.78	0.00	0.00	**72.70**	**48.86**	**73.72**	**42.38**	**38.40**	56.20
Qwen-VL-Chat	20.00	72.66	**56.39**	3.40	44.30	0.00	0.00	27.01	**48.85**	**40.05**	26.93	**20.40**	**63.20**
MoE-LLAVA-Phi2-2.7B-4e-384	20.20	67.67	**63.55**	1.90	**64.62**	0.00	0.00	**52.59**	**48.86**	34.24	36.83	**20.00**	**62.60**
mPLUG-Owl2-LLaMA2-7b	21.09	75.76	**60.74**	0.00	52.00	0.00	0.00	**47.87**	**47.60**	**40.00**	18.02	**21.20**	45.60
Phi-3.5-Vision-Instruct	20.40	**83.97**	**61.60**	3.44	**68.31**	0.00	0.00	**50.86**	**59.54**	**54.34**	25.74	**26.80**	38.40
Cambrian-1-8B	11.10	75.73	48.29	0.00	55.08	0.00	0.00	38.51	0.42	**43.00**	33.66	**19.60**	5.60
MiniGPT4-LLaMA2-7B	**28.80**	58.70	27.52	0.40	**65.23**	0.00	0.00	**51.44**	**82.65**	**77.18**	**97.32**	10.74	55.03
InternVL-Chat-V1-5	**45.80**	**86.90**	**81.40**	0.00	**79.60**	0.00	0.00	**64.30**	**83.60**	**56.00**	34.60	**32.60**	49.60
Mini-InternVL-Chat-4B-V1-5	**34.40**	81.10	**53.40**	0.00	**72.60**	0.00	0.00	**64.60**	**65.70**	**58.20**	27.90	**27.00**	56.00
InternLM-XComposer2-VL-1.8B	20.60	79.90	43.40	3.40	**75.00**	0.00	0.00	**42.50**	**78.90**	**67.50**	21.30	**21.60**	33.40
GPT4RoI	21.70	57.20	35.80	8.10	**69.80**	0.00	0.00	**51.10**	24.00	14.30	3.70	**16.70**	51.60
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	**42.60**	80.73	**66.80**	8.67	**73.85**	0.00	0.00	**52.87**	**51.97**	**47.21**	**42.18**	**35.60**	44.80
LLaVA-NeXT-34B	**52.60**	**83.90**	**70.40**	10.32	**76.00**	0.00	0.00	**60.35**	**56.39**	**43.44**	**39.80**	**37.40**	55.20
Pixtral 12B	**46.40**	77.43	**65.20**	3.25	**65.85**	0.00	0.00	**54.02**	**46.07**	**52.93**	34.06	**31.80**	49.20
SEED-LLaMA-13B	**28.30**	57.93	**53.80**	0.00	**63.69**	0.00	0.00	34.48	**46.68**	**55.87**	30.69	**28.20**	47.60
BLIP2	**33.90**	71.47	44.60	0.00	51.20	0.00	0.00	**46.26**	**40.91**	19.13	30.10	**25.00**	42.80
MiniMonkey	4.90	34.77	49.40	0.00	12.62	0.00	0.00	0.00	0.00	0.00	0.20	4.08	**69.20**
DeepSeek-VL-7B	**23.30**	79.53	**57.80**	0.00	**67.69**	0.00	0.00	**42.66**	**95.45**	**37.50**	32.28	**26.20**	**64.60**
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	**46.35**	**85.57**	**64.40**	12.32	**84.62**	0.00	0.00	**54.31**	**58.48**	**52.51**	**42.38**	**36.80**	54.20
ShareGPT4V-7B	**40.76**	78.20	**54.20**	8.45	**72.92**	0.00	0.00	38.79	**49.51**	**48.32**	30.50	**31.40**	48.40
ShareGPT4V-13B	**44.19**	80.03	**57.60**	10.37	**75.69**	0.00	0.00	**49.43**	**53.69**	**51.96**	35.45	**35.60**	52.60
BLIP-3 (XGen-MM)	**43.28**	**82.63**	**61.80**	8.54	**82.15**	0.00	0.00	**52.87**	**59.71**	**55.45**	37.43	**34.20**	50.60
AnyGPT	12.20	45.70	47.60	0.00	53.54	0.00	0.00	17.15	**43.98**	**47.17**	20.59	11.20	18.00
MiniCPM3-4B	**47.30**	79.13	**68.80**	4.68	**74.77**	0.00	0.00	**57.18**	**56.76**	**49.16**	**39.80**	**36.80**	49.20
LaVIT-V2 (7B)	**36.90**	55.17	**55.40**	0.00	62.15	0.00	0.00	34.48	**49.88**	**52.23**	32.87	**27.40**	48.60
GLM-VL-Chat	**51.10**	72.20	**70.20**	17.92	**81.85**	0.00	0.00	**63.79**	**53.44**	**56.28**	**44.16**	**39.80**	54.80
Gemini-1.5-Pro	**47.60**	**91.20**	**78.20**	23.41	**78.77**	0.00	0.00	**69.54**	**98.90**	**54.89**	**44.55**	**46.20**	60.20
Gemini-1.5-Flash	**42.10**	**87.30**	**71.60**	25.79	**75.69**	0.00	0.00	**67.24**	**98.03**	**52.65**	**40.20**	**42.20**	56.40
OMG-LLaVA-InternLM20B	3.10	4.20	1.40	4.03	3.70	0.00	0.00	6.90	7.25	3.77	1.58	2.60	3.40
Idefics3-8B-Llama3	**43.10**	79.30	**62.80**	11.50	**72.00**	0.00	0.00	**58.62**	**69.41**	**50.98**	**42.19**	**41.80**	52.20
NExT-GPT-V1.5	**34.89**	66.38	39.87	3.70	**65.34**	0.00	0.00	36.45	**43.79**	**53.47**	21.56	15.40	32.50
Vitron-V1	**39.78**	60.75	42.39	3.90	**66.75**	**36.40**	2.30	29.38	**50.32**	**53.46**	32.42	13.68	34.70
Otter	0.00	7.70	31.26	0.00	0.00	0.00	0.00	18.15	0.00	0.00	0.00	1.00	0.20
Show-o	20.00	0.18	4.21	0.00	0.00	0.00	0.00	31.89	**51.14**	**54.33**	20.19	2.21	37.40
NExT-Chat	19.92	45.76	**63.54**	0.00	**68.12**	0.00	0.00	26.82	**42.61**	**47.31**	11.74	**22.41**	16.64
Yi-vision-v2	**24.60**	79.40	**82.96**	2.13	**76.31**	0.00	0.00	**47.13**	**52.95**	15.78	**40.79**	**30.20**	60.00
Qwen2-VL-72B	**24.13**	**84.73**	**92.08**	5.74	**71.27**	0.00	0.00	**80.79**	**46.72**	**53.24**	**55.30**	**41.00**	62.40

Table 21: Results on Image Comprehension Group,#I-C-6, part A.

Model	#I-C-6 (Doc VQA)										
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑	#7 ↑	#8 ↑	#9 ↑	#10 ↑	#11 ↑
SoTA Specialist	32.72	10.64	11.76	19.40	17.80	26.71	20.92	30.94	16.11	23.91	13.79
GPT4V	**45.23**	**40.78**	**35.64**	**38.80**	**46.46**	**35.61**	**28.61**	27.61	**30.61**	**31.62**	**28.61**
GPT4o	**48.16**	**48.58**	**41.58**	**42.63**	**50.51**	**39.04**	**29.04**	30.04	**31.04**	**32.04**	**31.23**
GPT4o-mini	**41.18**	**38.30**	**40.59**	**38.80**	**45.45**	**30.83**	**30.83**	**31.83**	**32.83**	**33.83**	**30.23**
GPT-4o-4096	**49.63**	**47.52**	**46.53**	**46.45**	**51.18**	**45.21**	**43.51**	**46.41**	**48.59**	**42.81**	**46.29**
ChatGPT-4o-latest	**44.49**	**45.39**	**44.55**	**40.44**	**47.81**	**32.88**	**39.75**	**43.65**	**30.27**	**30.28**	**45.85**
Claude-3.5-Sonnet	**44.38**	**42.54**	**39.07**	**39.97**	**46.59**	**34.58**	**28.74**	28.91	**31.14**	**32.49**	**29.86**
Claude-3.5-Opus	**42.65**	**42.22**	**38.18**	**36.83**	**44.33**	**34.28**	**26.74**	26.22	**30.67**	**27.91**	**26.64**
Emu2-32B	32.72	**29.79**	**28.67**	**30.42**	**26.27**	25.25	19.06	21.40	**18.26**	19.45	**20.55**
DetGPT	30.88	**18.44**	**31.86**	**34.55**	**29.40**	**27.42**	19.47	22.87	**18.84**	19.64	**20.93**
InternVL2.5-8B	**33.45**	**28.47**	**43.99**	**41.89**	**35.77**	**30.83**	**35.06**	**36.67**	**41.67**	**25.49**	**35.23**
InternVL2.5-4B	31.54	**26.59**	**36.50**	**37.34**	**30.33**	**27.89**	**32.84**	**35.68**	**38.71**	**25.03**	**33.90**
InternVL2.5-2B	31.63	**29.07**	**32.56**	**37.77**	**33.72**	**29.91**	**32.53**	**33.15**	**40.21**	22.61	**33.48**
Monkey-10B-chat	32.42	**27.54**	**34.96**	**29.52**	**29.71**	**29.47**	**26.77**	30.89	**35.36**	**24.40**	**30.57**
DeepSeek-VL-7B-Chat	7.43	7.55	7.40	8.82	8.94	10.65	9.72	8.08	11.39	11.09	10.51
Qwen2-VL-7B	**45.83**	**52.00**	**46.70**	**62.66**	**50.00**	**46.66**	**52.94**	**68.18**	**51.06**	**36.73**	**58.62**
Qwen-VL-Chat	18.75	**16.66**	**23.76**	**19.67**	15.48	9.58	15.48	17.12	**17.39**	15.02	**15.72**
MoE-LLAVA-Phi2-2.7B-4e-384	11.40	9.57	**12.87**	8.19	9.76	12.32	11.72	10.77	9.72	10.99	8.73
mPLUG-Owl2-LLaMA2-7b	13.24	9.57	**18.81**	11.48	11.45	10.96	10.04	11.60	12.28	13.41	7.86
Phi-3.5-Vision-Instruct	**34.00**	**42.00**	**40.51**	**40.47**	**40.00**	**38.23**	**40.11**	**52.00**	**35.35**	**35.35**	**35.00**
Cambrian-1-8B	30.95	**31.91**	**29.41**	**21.21**	**38.57**	18.60	19.83	26.32	**30.16**	23.73	**31.03**
MiniGPT4-LLaMA2-7B	14.09	**17.45**	**15.00**	17.45	**18.79**	26.21	20.81	16.11	**20.81**	11.41	**16.78**
InternVL-Chat-V1-5	18.00	**20.20**	**13.80**	**21.70**	**60.00**	**33.30**	**22.20**	15.10	**57.10**	18.90	**15.70**
Mini-InternVL-Chat-4B-V1-5	32.70	**26.50**	**30.60**	**19.60**	**26.20**	21.20	**25.60**	22.90	**34.00**	**26.90**	13.70
InternLM-XComposer2-VL-1.8B	15.80	**19.80**	**14.80**	18.00	15.10	13.00	20.00	16.20	**24.80**	17.40	**16.50**
GPT4RoI	3.30	2.80	3.90	4.30	2.00	3.40	5.00	1.30	5.80	5.10	3.90
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	28.68	**28.01**	**38.61**	**37.30**	**33.67**	**35.62**	**24.69**	28.45	**26.85**	**26.33**	**27.07**
LLaVA-NeXT-34B	31.99	**34.04**	**42.57**	**42.70**	**38.05**	**36.99**	**28.45**	29.83	**24.81**	**28.92**	**28.82**
Pixtral 12B	**33.46**	**35.46**	**34.65**	**40.00**	**35.02**	**29.45**	**22.59**	25.69	**25.58**	**24.07**	**24.02**
SEED-LLaMA-13B	18.38	**19.15**	**14.85**	8.74	7.41	5.48	12.97	10.50	11.76	13.09	**14.41**
BLIP2	4.78	4.26	7.92	9.84	5.72	6.16	7.11	3.87	4.09	5.98	2.18
MiniMonkey	**34.96**	**31.22**	**32.67**	**36.93**	**33.67**	19.17	**30.54**	27.90	**33.73**	21.64	**25.76**
DeepSeek-VL-7B	31.99	**25.89**	**33.66**	**28.96**	**30.64**	22.60	**25.10**	30.11	**36.06**	21.00	**26.20**
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	30.88	**34.40**	**37.62**	**42.62**	**36.03**	**34.93**	**25.52**	29.56	**26.60**	**25.36**	**27.95**
ShareGPT4V-7B	22.43	**24.11**	**28.71**	**33.88**	**29.29**	26.71	**21.33**	23.48	**20.20**	20.03	**20.52**
ShareGPT4V-13B	25.37	**25.53**	**32.67**	**37.16**	**31.99**	**29.45**	**22.59**	25.41	**21.99**	21.65	**22.71**
BLIP-3 (XGen-MM)	28.68	**29.79**	**33.66**	**39.34**	**32.66**	**32.19**	**24.27**	28.18	**24.81**	23.10	**25.33**
AnyGPT	4.78	10.64	6.93	0.00	0.00	0.00	5.44	3.59	2.30	2.26	6.11
MiniCPM3-4B	**34.19**	**36.88**	**41.58**	**47.54**	**36.61**	**35.62**	**27.20**	30.11	**29.16**	21.81	**23.14**
LaVIT-V2 (7B)	16.18	**18.79**	**12.87**	12.02	9.29	0.00	11.28	10.50	11.00	12.12	**15.28**
GLM-VL-Chat	**35.66**	**38.65**	**42.57**	**22.95**	**39.34**	**39.04**	**29.71**	25.97	**27.37**	**27.14**	**29.26**
Gemini-1.5-Pro	**39.71**	**31.21**	**42.57**	**34.97**	**34.01**	**33.56**	**33.89**	30.39	**33.50**	**28.27**	**29.26**
Gemini-1.5-Flash	**35.29**	**26.24**	**35.64**	**26.23**	**29.29**	**28.77**	**26.78**	**31.49**	**29.16**	**27.14**	**25.76**
OMG-LLaVA-InternLM20B	1.47	1.42	1.98	1.64	1.35	2.05	1.67	1.66	1.53	1.78	2.18
Idefics3-8B-Llama3	29.41	**31.91**	**36.63**	**30.60**	**29.97**	**28.77**	**30.13**	26.52	**28.39**	**24.23**	**26.20**
NExT-GPT-V1.5	3.30	3.80	4.29	5.69	2.36	4.28	6.10	1.60	7.50	6.90	3.70
Vitron-V1	3.56	4.29	8.10	7.59	3.21	4.76	8.36	4.39	3.78	6.52	2.16
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	6.25	7.80	6.93	5.46	4.71	5.48	3.77	5.80	5.37	5.98	3.06
NExT-Chat	13.29	**14.55**	**13.04**	12.92	15.41	16.72	16.87	11.33	11.48	18.04	13.63
Yi-vision-v2	**40.32**	**31.03**	**35.48**	**29.73**	**37.50**	**38.24**	**31.25**	**38.78**	**42.62**	**34.15**	**35.71**
Qwen2-VL-72B	29.86	**37.06**	**18.37**	**54.63**	**30.36**	**43.27**	**33.74**	**55.22**	**54.27**	**42.94**	**36.17**

Table 22: Results on Image Comprehension Group,#I-C-6, part B.

Model	#I-C-6 (Doc VQA)									
	#12 ↑	#13 ↑	#14 ↑	#15 ↑	#16 ↑	#17 ↑	#18 ↑	#19 ↑	#20 ↑	#21 ↑
SoTA Specialist	11.76	21.51	21.51	24.06	14.02	11.64	23.26	16.67	19.51	82.20
GPT4V	**32.35**	**53.70**	**35.70**	**34.10**	**25.90**	**14.30**	**51.50**	**41.20**	**36.20**	43.21
GPT4o	**33.82**	**50.90**	**44.00**	**35.70**	**26.10**	**43.80**	**48.70**	**42.20**	**34.30**	49.16
GPT4o-mini	**29.42**	**40.40**	**42.00**	**30.90**	**33.30**	**25.00**	**39.60**	**40.20**	**31.60**	40.08
GPT-4o-4096	**38.24**	**48.11**	**44.37**	**45.94**	**42.58**	**44.60**	**53.49**	**46.67**	**47.76**	**95.90**
ChatGPT-4o-latest	**30.88**	**47.03**	**40.40**	**40.94**	**16.13**	**40.30**	**51.62**	**42.38**	**35.26**	48.28
Claude-3.5-Sonnet	**31.41**	**48.16**	**39.76**	**32.75**	**27.96**	**27.67**	**46.05**	**40.66**	**33.43**	43.90
Claude-3.5-Opus	**28.32**	**44.09**	**37.23**	**30.11**	**24.71**	**24.46**	**46.05**	**39.79**	**30.15**	42.83
Emu2-32B	**19.82**	**24.25**	8.96	21.35	**16.85**	**19.97**	**24.98**	**23.45**	**21.22**	14.03
DetGPT	**20.46**	**26.42**	5.82	21.88	**16.75**	**20.36**	**26.60**	**24.99**	**22.42**	12.14
InternVL2.5-8B	**38.78**	**27.03**	**34.99**	**40.85**	**23.50**	**26.54**	**45.26**	**41.02**	**38.20**	38.41
InternVL2.5-4B	**34.27**	**25.32**	**34.40**	**37.95**	**20.95**	**25.62**	**49.59**	**38.43**	**35.43**	31.93
InternVL2.5-2B	**32.44**	**23.02**	**34.67**	**40.45**	**23.62**	**24.74**	**47.28**	**39.11**	**31.78**	33.33
Monkey-10B-chat	**25.43**	**27.47**	**30.88**	**36.00**	**22.75**	**26.32**	**38.19**	**32.31**	**35.43**	24.84
DeepSeek-VL-7B-Chat	7.96	8.85	9.96	14.60	10.71	8.59	15.29	11.17	8.66	8.02
Qwen2-VL-7B	**45.61**	**43.90**	**43.50**	**46.42**	**30.55**	**38.00**	**32.52**	**44.47**	**46.00**	45.61
Qwen-VL-Chat	**18.37**	15.23	17.50	16.17	13.54	**14.35**	18.60	14.28	16.32	15.06
MoE-LLAVA-Phi2-2.7B-4e-384	11.76	11.89	8.61	12.81	9.03	10.03	11.63	6.19	7.76	42.80
mPLUG-Owl2-LLaMA2-7b	11.35	10.60	8.75	5.88	10.65	11.42	13.95	6.19	11.02	38.94
Phi-3.5-Vision-Instruct	**47.90**	**44.23**	**43.00**	**41.30**	**38.25**	**33.70**	**55.11**	**50.00**	**47.36**	49.00
Cambrian-1-8B	**25.00**	**38.46**	11.54	21.18	**20.00**	**22.73**	17.65	**17.50**	**26.83**	43.46
MiniGPT4-LLaMA2-7B	**21.48**	18.79	**28.08**	16.11	**14.09**	**22.82**	**33.33**	**22.82**	16.78	2.72
InternVL-Chat-V1-5	**20.00**	**33.30**	17.20	**41.00**	**22.90**	**17.50**	13.90	**46.10**	**25.00**	44.10
Mini-InternVL-Chat-4B-V1-5	**32.30**	**22.10**	**24.50**	**33.70**	13.70	**23.40**	**34.80**	**33.80**	**24.00**	41.50
InternLM-XComposer2-VL-1.8B	**22.00**	14.50	21.10	20.90	**17.00**	**15.50**	23.20	13.30	13.40	39.10
GPT4RoI	4.40	4.80	3.90	3.40	7.00	3.80	9.30	1.00	2.40	1.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	**23.53**	**32.43**	18.60	**27.81**	**20.97**	**25.15**	**27.91**	**30.00**	**25.71**	32.40
LLaVA-NeXT-34B	**30.88**	**35.14**	**23.50**	**30.94**	**24.52**	**27.62**	**32.56**	**35.71**	**31.84**	36.20
Pixtral 12B	**20.59**	**28.65**	12.40	**25.94**	**23.55**	**19.29**	20.93	**32.38**	**28.57**	30.80
SEED-LLaMA-13B	**17.65**	15.14	4.80	13.13	**14.19**	9.57	**30.23**	7.62	18.37	15.60
BLIP2	2.94	1.99	6.45	5.31	9.68	2.93	9.30	0.95	1.22	16.59
MiniMonkey	**27.94**	**33.11**	**23.93**	**36.62**	**16.77**	**27.47**	**34.88**	**30.00**	**36.73**	20.63
DeepSeek-VL-7B	**25.00**	**25.83**	**27.53**	**34.06**	**22.58**	**25.77**	**34.88**	**31.43**	**34.29**	8.00
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	**25.00**	**34.59**	21.19	**27.81**	**22.90**	**27.16**	**37.21**	**30.48**	**27.35**	31.50
ShareGPT4V-7B	**17.65**	**27.57**	13.91	22.19	**18.39**	**22.07**	23.26	**24.76**	**20.82**	24.94
ShareGPT4V-13B	**20.59**	**29.19**	17.22	23.75	**19.68**	**24.23**	**27.91**	**26.67**	**23.27**	26.19
BLIP-3 (XGen-MM)	**26.47**	**31.35**	19.20	**25.31**	**21.61**	**25.31**	**32.56**	**29.05**	**26.12**	28.47
AnyGPT	8.82	5.95	5.96	1.88	4.94	2.16	0.47	0.95	5.71	0.00
MiniCPM3-4B	**29.41**	**37.84**	11.26	**30.31**	**24.84**	**16.05**	**32.56**	**34.29**	**32.65**	35.64
LaVIT-V2 (7B)	**16.18**	20.55	9.27	11.25	12.58	9.10	20.93	6.67	17.55	14.77
GLM-VL-Chat	**27.94**	**36.76**	19.21	**32.50**	**24.19**	**33.02**	**44.19**	**30.95**	**31.02**	32.17
Gemini-1.5-Pro	**32.35**	**42.16**	**35.76**	**31.25**	**29.03**	**25.31**	**39.60**	**38.57**	**34.29**	40.60
Gemini-1.5-Flash	**29.42**	**38.92**	**29.80**	**29.38**	**24.19**	**19.75**	**34.88**	**35.24**	**28.57**	36.20
OMG-LLaVA-InternLM20B	2.94	3.24	2.65	2.50	1.61	1.54	0.00	2.38	1.63	0.00
Idefics3-8B-Llama3	**26.47**	**33.11**	**34.40**	**28.13**	**30.00**	**30.56**	**32.56**	**38.57**	**31.43**	29.80
NExT-GPT-V1.5	3.60	5.78	2.63	4.78	7.60	0.36	5.48	1.08	3.75	1.00
Vitron-V1	3.03	2.39	0.24	4.75	8.31	3.96	10.76	1.39	5.68	3.70
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	4.41	8.11	3.97	4.06	3.23	5.40	4.65	4.76	6.94	0.00
NExT-Chat	**17.83**	16.54	13.70	19.14	12.24	0.00	17.45	11.59	11.93	15.74
Yi-vision-v2	**30.77**	**43.34**	**36.26**	**32.39**	**30.30**	**25.93**	9.30	**39.24**	**34.38**	7.38
Qwen2-VL-72B	**44.61**	**25.64**	**56.50**	**48.99**	**40.27**	**57.94**	**37.07**	**25.15**	**28.00**	50.43

Table 23: Results on Image Comprehension Group, from #I-C-7 to #I-C-10.

Model	#I-C-7 (Emotion Det)				#I-C-8 (Graph Cls)	#I-C-9 (Hallucination Det)		#I-C-10 (Img Cap)					
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#1 ↑	#1 ↑	#2 ↑	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑
SoTA Specialist	31.12	59.44	41.20	52.30	15.67	41.60	60.80	17.84	62.99	23.05	18.40	17.65	28.10
GPT4V	**51.06**	**61.45**	**53.83**	**80.60**	0.00	12.80	21.00	2.63	42.22	15.80	**18.70**	**27.21**	18.69
GPT4o	**63.65**	**63.86**	**61.51**	**86.30**	0.00	**69.60**	**66.00**	1.30	48.10	16.51	**19.81**	**30.42**	23.30
GPT4o-mini	**61.38**	56.62	**64.57**	**84.30**	0.00	40.80	**78.20**	4.00	43.10	18.64	15.05	**25.62**	18.60
GPT-4o-4096	**65.63**	**64.66**	**66.39**	**88.70**	0.00	**68.20**	**70.80**	1.00	45.45	17.36	**20.68**	**28.86**	15.72
ChatGPT-4o-latest	**55.02**	**62.25**	**59.41**	**81.11**	0.00	**50.00**	**64.20**	1.74	41.23	15.82	15.85	**21.03**	13.53
Claude-3.5-Sonnet	**58.31**	**59.89**	**59.92**	**83.26**	0.00	40.40	54.25	1.96	44.25	16.80	17.63	**26.82**	19.57
Claude-3.5-Opus	**57.71**	56.41	**57.87**	**81.74**	0.00	38.13	53.11	0.74	41.23	16.02	17.73	**26.82**	20.09
Emu2-32B	**32.67**	40.76	**45.61**	48.51	0.00	17.60	24.80	0.54	37.13	12.42	11.34	0.00	15.53
DetGPT	21.64	32.93	**41.56**	**58.69**	0.00	6.80	5.20	0.19	29.04	8.65	6.22	0.00	13.86
InternVL2.5-8B	**45.54**	**65.06**	**47.14**	**72.41**	0.00	4.29	11.23	9.06	16.99	20.42	13.79	7.49	7.02
InternVL2.5-4B	**47.10**	**61.45**	**45.75**	**100.00**	0.00	13.13	32.62	8.24	16.43	20.53	13.37	14.59	4.61
InternVL2.5-2B	**49.79**	58.23	**50.21**	**67.22**	0.00	3.79	9.41	9.58	16.03	20.62	7.47	7.83	7.36
Monkey-10B-chat	**36.78**	46.99	20.22	**57.78**	0.00	5.93	43.69	13.83	2.45	17.89	0.00	0.00	6.35
DeepSeek-VL-7B-Chat	**47.67**	**62.65**	**48.50**	**78.52**	0.00	3.48	3.18	3.10	12.75	17.77	17.84	**19.03**	3.64
Qwen2-VL-7B	29.00	**61.45**	28.73	**73.04**	0.00	**52.20**	24.91	4.40	57.14	**26.50**	**20.60**	**36.62**	13.77
Qwen-VL-Chat	**32.67**	46.98	25.52	**64.62**	0.00	2.80	36.19	4.00	58.61	**26.28**	**20.66**	**24.85**	15.33
MoE-LLAVA-Phi2-2.7B-4e-384	**50.35**	56.63	**48.26**	48.28	0.00	0.00	7.60	4.11	57.39	**28.36**	**18.74**	13.29	10.79
mPLUG-Owl2-LLaMA2-7b	**42.29**	54.62	39.03	40.17	0.00	0.00	1.20	4.07	52.69	**26.30**	**20.00**	11.91	10.32
Phi-3.5-Vision-Instruct	19.31	**60.64**	**46.72**	**78.70**	0.00	0.00	0.00	4.63	54.51	**26.84**	**24.50**	**20.20**	15.40
Cambrian-1-8B	14.29	49.80	**44.70**	51.40	0.00	1.40	1.00	4.20	30.60	19.92	10.44	**21.50**	10.09
MiniGPT4-LLaMA2-7B	**33.66**	40.56	**74.50**	**87.25**	0.00	**46.98**	55.70	4.47	37.40	**26.54**	16.44	0.00	4.54
InternVL-Chat-V1-5	**49.30**	55.80	**49.20**	**76.20**	0.00	38.20	46.40	4.30	20.40	14.20	7.60	4.50	7.90
Mini-InternVL-Chat-4B-V1-5	**45.60**	41.30	31.50	**65.70**	0.00	25.80	40.60	4.60	19.90	13.50	3.70	2.70	1.90
InternLM-XComposer2-VL-1.8B	29.30	39.30	**47.60**	**62.70**	0.00	7.80	48.20	4.20	25.40	12.20	2.80	1.40	11.30
GPT4RoI	**47.80**	46.10	40.40	43.30	0.00	4.20	29.80	3.90	18.40	12.70	2.50	0.00	6.80
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.80	23.60	13.30	2.90	0.00	7.00
LLaVA-NeXT-13B	**36.21**	48.59	34.73	**62.41**	0.00	15.20	18.40	1.18	32.48	22.05	14.32	0.00	18.12
LLaVA-NeXT-34B	**45.69**	55.42	**47.00**	**71.48**	0.00	18.80	25.60	1.04	36.12	**23.45**	15.96	0.00	17.49
Pixtral 12B	**38.47**	47.38	39.05	**64.07**	0.00	16.40	21.20	0.58	30.56	19.85	12.86	0.00	18.01
SEED-LLaMA-13B	26.73	35.74	34.03	43.89	0.00	5.80	8.60	1.12	24.12	9.84	7.56	0.00	11.92
BLIP2	23.06	43.37	29.43	51.11	0.00	18.20	20.40	6.74	15.24	17.26	11.30	0.00	8.30
MiniMonkey	12.16	21.69	11.16	**58.63**	0.00	4.60	29.80	0.31	29.98	18.91	0.00	0.00	4.42
DeepSeek-VL-7B	26.59	58.78	**44.21**	**73.26**	0.00	5.80	6.20	2.02	18.74	19.04	12.60	0.00	4.06
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	**44.55**	48.59	**45.19**	**69.63**	0.00	17.40	23.60	2.56	38.74	22.04	16.47	0.00	23.15
ShareGPT4V-7B	30.13	39.36	30.26	**56.85**	0.00	14.80	19.20	1.48	33.92	**24.46**	11.59	0.00	16.72
ShareGPT4V-13B	**31.26**	45.78	31.24	**60.37**	0.00	15.60	21.60	1.85	37.30	**25.08**	13.95	0.00	18.38
BLIP-3 (XGen-MM)	**42.57**	42.17	**43.79**	**63.33**	0.00	16.20	22.40	2.17	29.77	**24.45**	14.36	0.00	21.79
AnyGPT	23.62	26.03	19.94	34.26	0.00	0.00	0.00	0.42	23.16	16.24	6.63	0.00	11.69
MiniCPM3-4B	**48.37**	52.21	**53.56**	**64.26**	0.00	18.40	22.40	1.24	39.13	20.04	14.91	**28.93**	22.67
LaVIT-V2 (7B)	**37.06**	44.18	34.59	46.11	0.00	4.40	14.20	0.91	29.28	12.08	10.08	0.00	16.45
GLM-VL-Chat	**46.82**	55.02	**50.49**	**74.63**	0.00	19.80	26.20	1.97	43.50	**23.84**	16.30	**27.64**	27.98
Gemini-1.5-Pro	**58.58**	**69.08**	**55.65**	**81.67**	0.00	**53.80**	**64.60**	4.49	44.50	22.34	**19.40**	**30.73**	21.30
Gemini-1.5-Flash	**57.85**	**67.07**	**52.86**	**76.67**	0.00	**52.00**	**61.00**	4.51	39.72	**23.58**	**18.95**	**28.58**	17.80
OMG-LLaVA-InternLM20B	6.36	13.65	3.49	3.89	0.00	2.20	1.60	3.24	22.37	18.28	11.37	0.00	7.50
Idefics3-8B-Llama3	**45.83**	51.41	35.70	**71.30**	0.00	28.80	41.20	2.60	42.55	22.76	13.85	**22.33**	13.10
NExT-GPT-V1.5	28.90	33.57	37.56	46.10	0.00	12.70	10.67	2.65	23.70	19.78	11.57	12.39	2.10
Vitron-V1	**33.40**	34.10	38.16	**53.47**	0.00	16.20	10.40	2.36	35.10	21.86	10.39	10.56	2.60
Otter	14.59	10.89	3.91	3.33	7.17	1.33	1.40	0.56	4.26	13.45	0.00	0.00	0.00
Show-o	10.34	12.61	17.15	17.22	0.00	0.00	0.00	0.00	38.07	11.50	0.00	0.00	0.00
NExT-Chat	25.43	**70.63**	0.00	27.39	0.00	0.00	0.00	3.49	56.19	12.28	15.13	0.00	0.00
Yi-vision-v2	21.07	53.41	16.60	35.48	0.00	**66.93**	52.79	4.24	22.80	15.02	12.47	0.00	13.49
Qwen2-VL-72B	28.22	**61.88**	26.47	**78.57**	0.00	**54.31**	32.04	4.79	57.37	**27.02**	16.48	**29.30**	16.98

Table 24: Results on Image Comprehension Group, from #I-C-11 to #I-C-14.

Model	#I-C-11 (Img Depth Est)	#I-C-12 (Img Inst Seg)		#I-C-13 (Img OCR)				#I-C-14 (Img Recog)	
	#1 ↓	#1 ↑	#2 ↑	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#1 ↑	#2 ↑
SoTA Specialist	36.40	66.50	63.80	45.03	95.14	10.23	24.70	29.80	88.00
GPT4V	∞	0.00	0.00	**89.71**	35.68	**60.45**	18.30	**32.50**	**94.53**
GPT4o	∞	0.00	0.00	**94.07**	**97.36**	**74.01**	19.46	27.83	**95.24**
GPT4o-mini	∞	0.00	0.00	**85.10**	**96.56**	**73.12**	17.42	23.83	**91.32**
GPT-4o-4096	∞	0.00	0.00	**94.22**	**96.67**	**74.45**	0.00	**34.00**	32.20
ChatGPT-4o-latest	∞	0.00	0.00	**93.36**	94.57	**73.99**	18.43	**30.23**	**93.23**
Claude-3.5-Sonnet	∞	0.00	0.00	**89.06**	75.64	**68.20**	17.55	28.04	**93.26**
Claude-3.5-Opus	∞	0.00	0.00	**87.69**	71.88	**64.89**	16.36	26.57	**89.73**
Emu2-32B	∞	0.00	0.00	39.75	61.39	**49.12**	7.60	24.40	0.00
DetGPT	∞	0.00	0.00	42.13	64.21	**56.19**	3.10	13.17	0.00
InternVL2.5-8B	∞	0.00	0.00	27.46	43.70	**35.03**	0.54	24.17	11.31
InternVL2.5-4B	∞	0.00	0.00	24.31	41.53	**29.27**	0.00	14.83	12.62
InternVL2.5-2B	∞	0.00	0.00	27.39	31.06	**43.47**	0.00	12.83	9.34
Monkey-10B-chat	∞	0.00	0.00	0.00	0.00	0.00	0.00	**37.50**	1.15
DeepSeek-VL-7B-Chat	∞	0.00	0.00	22.98	45.90	**14.11**	0.27	**33.33**	4.38
Qwen2-VL-7B	∞	0.00	0.00	16.70	90.78	**24.72**	4.41	**30.47**	11.98
Qwen-VL-Chat	∞	0.00	0.00	20.10	91.52	**19.63**	4.06	**33.50**	9.00
MoE-LLAVA-Phi2-2.7B-4e-384	∞	0.00	0.00	22.28	93.10	**20.15**	0.40	**33.50**	6.24
mPLUG-Owl2-LLaMA2-7b	∞	0.00	0.00	18.89	84.25	**12.90**	0.00	**33.83**	3.50
Phi-3.5-Vision-Instruct	∞	0.00	0.00	41.67	92.60	**26.76**	2.97	**32.67**	10.87
Cambrian-1-8B	∞	0.00	0.00	16.77	84.19	**20.80**	0.00	14.00	3.81
MiniGPT4-LLaMA2-7B	∞	0.00	0.00	22.50	83.35	**16.54**	0.31	**39.50**	4.37
InternVL-Chat-V1-5	∞	0.00	0.00	42.60	92.50	**40.00**	0.00	**69.30**	25.10
Mini-InternVL-Chat-4B-V1-5	∞	0.00	0.00	27.80	61.30	**38.80**	0.00	**59.80**	18.50
InternLM-XComposer2-VL-1.8B	∞	0.00	0.00	22.10	12.90	**31.70**	0.00	**31.00**	15.50
GPT4RoI	∞	0.00	0.00	0.00	0.00	0.00	0.00	22.60	14.10
GLaMM	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	∞	0.00	0.00	**62.38**	72.34	**62.57**	6.40	27.33	7.30
LLaVA-NeXT-34B	∞	0.00	0.00	**69.89**	78.05	**70.08**	8.20	28.67	14.40
Pixtral 12B	∞	0.00	0.00	42.34	64.34	**56.05**	4.80	24.33	5.80
SEED-LLaMA-13B	∞	0.00	0.00	22.03	32.57	**47.08**	0.00	16.00	0.00
BLIP2	∞	0.00	0.00	23.81	44.38	**46.92**	3.60	**38.00**	0.00
MiniMonkey	∞	0.00	0.00	0.00	0.00	0.00	0.00	**38.00**	2.84
DeepSeek-VL-7B	∞	0.00	0.00	**61.82**	47.09	**32.15**	0.35	**37.50**	5.67
LISA	∞	42.09	**92.75**	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	∞	0.00	0.00	**72.69**	86.53	**70.89**	8.35	27.00	38.20
ShareGPT4V-7B	∞	0.00	0.00	**54.21**	69.44	**56.47**	6.72	21.50	27.40
ShareGPT4V-13B	∞	0.00	0.00	**59.24**	72.31	**60.36**	7.55	22.83	32.80
BLIP-3 (XGen-MM)	∞	0.00	0.00	**68.57**	82.47	**63.52**	7.24	25.67	35.60
AnyGPT	∞	0.00	0.00	0.00	31.14	**19.76**	0.00	6.67	14.80
MiniCPM3-4B	∞	0.00	0.00	**71.96**	77.08	**71.85**	9.80	29.50	73.60
LaVIT-V2 (7B)	∞	0.00	0.00	43.16	57.98	**47.36**	0.00	15.17	32.40
GLM-VL-Chat	∞	0.00	0.00	**74.28**	61.42	**64.27**	8.40	25.17	69.00
Gemini-1.5-Pro	∞	0.00	0.00	**83.56**	89.73	**70.15**	0.00	25.83	54.36
Gemini-1.5-Flash	∞	0.00	0.00	**73.48**	82.81	**64.58**	0.00	23.67	42.17
OMG-LLaVA-InternLM20B	∞	39.50	**68.22**	18.91	14.63	8.92	0.00	2.00	8.95
Idefics3-8B-Llama3	∞	0.00	0.00	43.02	62.25	**59.32**	0.00	20.50	16.57
NExT-GPT-V1.5	∞	0.00	0.00	22.56	56.17	**45.68**	0.00	13.65	0.00
Vitron-V1	∞	**68.70**	**64.50**	24.51	63.01	**70.34**	0.00	16.37	0.00
Otter	∞	0.00	0.00	11.33	47.72	**17.00**	0.00	0.00	0.00
Show-o	∞	0.00	0.00	0.00	0.00	0.00	0.00	25.17	0.00
NExT-Chat	∞	0.00	0.00	**76.75**	71.32	**61.04**	0.00	17.41	0.00
Yi-vision-v2	∞	0.00	0.00	**83.57**	90.28	**55.23**	0.00	**35.67**	16.83
Qwen2-VL-72B	∞	0.00	0.00	26.58	**95.27**	**60.79**	0.00	**45.16**	13.72

Table 25: Results on Image Comprehension Group, from #I-C-15 to #I-C-16.

Model	#I-C-15 (Img Sem Seg)								#I-C-16 (Img Vis Grnd)	
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑	#7 ↑	#8 ↑	#1 ↑	#2 ↑
SoTA Specialist	19.80	50.40	88.00	98.70	83.65	59.36	47.74	62.15	84.76	90.91
GPT4V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	67.69	74.11
GPT4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	76.20	82.56
GPT4o-mini	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	57.18	59.44
GPT-4o-4096	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	80.60	83.33
ChatGPT-4o-latest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	68.30	70.70
Claude-3.5-Sonnet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.50	71.66
Claude-3.5-Opus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	62.40	70.73
Emu2-32B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DetGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	48.48	82.23
InternVL2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Monkey-10B-chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MoE-LLAVA-Phi2-2.7B-4e-384	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mPLUG-Owl2-LLaMA2-7b	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phi-3.5-Vision-Instruct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cambrian-1-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniGPT4-LLaMA2-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-Chat-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mini-InternVL-Chat-4B-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternLM-XComposer2-VL-1.8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	**42.10**	0.00	33.90	83.60	0.00	0.00	**57.40**	**64.20**	0.00	72.70
LLaVA-NeXT-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-34B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pixtral 12B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SEED-LLaMA-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniMonkey	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LISA	9.35	**90.28**	44.45	22.45	**88.74**	**88.65**	39.46	0.00	0.00	0.00
CogVLM-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP-3 (XGen-MM)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AnyGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniCPM3-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LaVIT-V2 (7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLM-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OMG-LLaVA-InternLM20B	**49.65**	42.57	67.73	20.15	79.05	40.26	44.80	56.49	0.00	0.00
Idefics3-8B-Llama3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	6.40	**52.40**	81.00	89.70	78.40	38.90	**56.70**	**64.70**	78.70	86.73
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Yi-vision-v2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 26: Results on Image Comprehension Group,#I-C-17, part A.

Model	#I-C-17 (Img VQA)									
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑	#7 ↑	#8 ↑	#9 ↑	#10 ↑
SoTA Specialist	14.36	36.40	17.32	43.12	89.38	17.50	34.30	80.75	92.72	78.30
GPT4V	**16.20**	26.54	**52.84**	41.04	80.38	**45.74**	**37.00**	57.33	87.54	43.33
GPT4o	**16.00**	23.82	**50.31**	31.41	85.75	**69.03**	**37.00**	53.00	**96.10**	51.00
GPT4o-mini	11.45	17.45	**47.24**	29.09	82.01	**44.03**	34.00	33.33	**93.10**	37.67
GPT-4o-4096	**24.40**	22.73	**53.78**	**43.16**	89.02	**74.15**	**40.00**	52.00	**96.63**	48.67
ChatGPT-4o-latest	**16.82**	19.64	**49.20**	**51.64**	83.88	**69.32**	**36.98**	63.33	84.81	41.67
Claude-3.5-Sonnet	13.98	21.73	**49.57**	33.44	81.97	**52.38**	**35.48**	47.78	91.38	43.34
Claude-3.5-Opus	13.40	20.90	**47.42**	28.99	77.92	**48.74**	**34.46**	43.76	90.51	42.52
Emu2-32B	10.18	34.55	**33.25**	38.54	65.42	**35.51**	31.67	43.64	65.32	33.67
DetGPT	9.36	29.09	**21.19**	28.90	55.84	**28.41**	25.00	34.55	63.97	28.33
InternVL2.5-8B	3.45	**47.27**	**31.54**	39.31	55.74	**60.00**	**70.67**	**83.64**	**98.65**	43.67
InternVL2.5-4B	10.64	**37.82**	**34.16**	40.66	27.02	**46.57**	**38.33**	**86.36**	**97.31**	44.33
InternVL2.5-2B	**29.09**	**40.55**	**38.97**	39.69	36.48	**49.14**	**84.00**	**88.18**	**93.10**	38.00
Monkey-10B-chat	6.91	3.60	**45.64**	8.11	14.92	**66.29**	**55.11**	**85.67**	72.05	26.00
DeepSeek-VL-7B-Chat	3.45	**100.00**	0.22	0.00	33.78	6.57	**66.67**	53.10	58.94	24.44
Qwen2-VL-7B	**54.91**	**65.45**	**54.32**	37.34	26.87	0.00	34.00	61.82	54.04	29.30
Qwen-VL-Chat	**57.99**	**52.90**	**38.96**	33.26	67.99	5.71	33.33	54.54	71.54	18.66
MoE-LLAVA-Phi2-2.7B-4e-384	**40.00**	**54.00**	**33.36**	31.44	67.29	7.80	33.67	60.28	66.16	14.67
mPLUG-Owl2-LLaMA2-7b	**38.00**	**46.73**	15.67	33.80	53.27	8.12	**39.00**	37.27	58.58	13.33
Phi-3.5-Vision-Instruct	**40.18**	**60.00**	**41.60**	35.90	78.04	**31.14**	**42.33**	**81.82**	**94.95**	32.67
Cambrian-1-8B	5.82	1.45	**29.00**	28.70	32.94	0.00	2.00	28.18	26.26	0.33
MiniGPT4-LLaMA2-7B	**34.23**	**38.26**	**46.83**	30.06	55.03	**32.43**	28.19	32.11	68.46	38.93
InternVL-Chat-V1-5	11.20	23.90	6.40	**47.90**	62.90	**22.40**	**34.60**	58.30	68.40	26.60
Mini-InternVL-Chat-4B-V1-5	9.80	16.70	4.50	**44.30**	59.20	16.32	11.33	51.70	62.90	20.00
InternLM-XComposer2-VL-1.8B	6.70	15.60	3.20	40.50	36.80	10.60	33.30	47.90	55.30	23.90
GPT4RoI	4.00	7.40	0.00	**52.80**	36.00	5.60	23.30	67.80	46.80	68.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	**14.55**	33.30	**36.18**	40.46	80.14	**41.76**	30.33	34.55	83.83	37.33
LLaVA-NeXT-34B	**15.09**	**36.80**	**40.25**	**48.17**	78.84	**52.27**	33.67	40.00	87.88	43.67
Pixtral 12B	12.00	23.50	**33.85**	42.78	71.96	**44.03**	31.00	32.73	74.75	36.00
SEED-LLaMA-13B	10.55	26.70	**28.96**	31.21	62.85	**37.78**	28.67	28.18	65.99	32.33
BLIP2	9.45	**46.00**	**27.64**	38.84	25.46	**29.26**	29.00	22.73	59.26	29.33
MiniMonkey	6.01	25.45	**39.31**	37.80	66.59	**45.92**	33.33	**84.54**	**93.94**	36.00
DeepSeek-VL-7B	12.78	31.25	**32.98**	35.63	58.18	**57.10**	**41.33**	51.82	56.06	23.67
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	11.45	**37.09**	**36.43**	**45.28**	81.31	**52.84**	**35.33**	47.27	81.82	44.67
ShareGPT4V-7B	7.82	25.82	**27.95**	39.11	71.03	**42.33**	31.33	34.55	75.93	35.67
ShareGPT4V-13B	9.09	30.36	**30.44**	**43.35**	76.40	**46.88**	33.67	41.82	79.97	38.67
BLIP-3 (XGen-MM)	9.45	35.64	**28.57**	42.58	73.36	**49.72**	32.00	44.55	77.95	41.00
AnyGPT	0.00	6.36	17.26	29.11	40.19	**21.02**	3.67	12.73	35.52	3.67
MiniCPM3-4B	12.18	31.45	**49.66**	**50.47**	79.91	**49.15**	31.67	46.37	89.23	40.33
LaVIT-V2 (7B)	9.45	20.36	**29.97**	42.91	51.87	**37.50**	15.00	28.19	64.98	32.67
GLM-VL-Chat	12.73	**37.64**	**41.20**	**51.98**	84.11	**57.67**	34.30	50.91	86.03	44.33
Gemini-1.5-Pro	14.18	18.73	**46.72**	**52.41**	86.68	**59.94**	**36.67**	**83.64**	**94.44**	48.33
Gemini-1.5-Flash	13.27	17.27	**43.80**	42.77	85.98	**54.83**	34.00	**82.73**	90.91	44.67
OMG-LLaVA-InternLM20B	0.00	2.55	**22.16**	2.12	4.91	1.42	1.67	2.73	2.19	2.67
Idefics3-8B-Llama3	12.91	16.73	**36.46**	34.30	67.99	**44.03**	28.67	57.27	82.15	38.33
NExT-GPT-V1.5	3.40	28.90	**31.54**	30.50	47.60	**34.96**	13.40	24.63	63.45	29.40
Vitron-V1	5.60	31.50	**34.78**	27.30	51.40	**31.84**	20.70	18.90	67.82	24.70
Otter	0.00	0.00	7.34	0.00	14.10	17.29	9.67	0.00	0.00	0.00
Show-o	**14.72**	**39.24**	**33.47**	0.00	0.00	0.85	**34.33**	1.82	0.51	0.33
NExT-Chat	7.41	34.83	**41.67**	0.00	38.86	11.63	**43.29**	41.67	40.37	37.93
Yi-vision-v2	14.00	**41.64**	**53.94**	0.00	83.41	**45.63**	**37.00**	**83.64**	**96.13**	35.67
Qwen2-VL-72B	**69.09**	**51.79**	**55.10**	0.00	0.00	**50.58**	**47.33**	59.00	73.82	49.42

Table 27: Results on Image Comprehension Group,#I-C-17, part B.

Model	#I-C-17 (Img VQA)									
	#11 ↑	#12 ↑	#13 ↑	#14 ↑	#15 ↑	#16 ↑	#17 ↑	#18 ↑	#19 ↑	#20 ↑
SoTA Specialist	49.80	65.40	66.50	74.70	73.58	24.40	71.60	70.70	51.60	71.40
GPT4V	0.00	50.00	**75.70**	**81.82**	**76.39**	**34.25**	58.70	66.00	47.33	70.40
GPT4o	0.00	**68.10**	**79.44**	72.72	**78.17**	**45.00**	66.30	68.33	12.70	**72.20**
GPT4o-mini	0.00	56.90	**71.96**	**77.27**	69.27	**28.00**	56.52	69.67	12.60	63.20
GPT-4o-4096	0.00	64.66	**83.18**	**81.82**	**79.29**	**45.25**	54.35	68.33	50.33	**75.60**
ChatGPT-4o-latest	0.00	62.07	**79.44**	**77.27**	**75.72**	**45.00**	57.61	70.33	47.00	**75.20**
Claude-3.5-Sonnet	0.00	57.81	**75.57**	**76.59**	**74.35**	**35.54**	60.07	67.93	24.13	50.60
Claude-3.5-Opus	0.00	54.04	**72.74**	72.86	71.10	**32.88**	59.55	64.81	19.68	53.73
Emu2-32B	**53.67**	47.41	65.42	54.55	66.82	**27.50**	55.43	54.00	29.67	32.80
DetGPT	34.33	37.07	56.07	31.82	54.34	**25.50**	43.48	47.33	27.33	34.40
InternVL2.5-8B	**72.00**	13.79	**69.16**	72.73	**87.31**	**41.50**	36.96	43.66	50.67	53.85
InternVL2.5-4B	47.67	42.24	**75.70**	**77.27**	**95.09**	**27.37**	71.20	67.72	41.22	**86.04**
InternVL2.5-2B	**99.33**	16.09	60.75	**77.27**	**83.96**	**62.50**	41.30	45.77	0.00	**87.04**
Monkey-10B-chat	**50.00**	43.10	57.94	54.55	54.34	**54.67**	57.61	49.30	49.33	**81.60**
DeepSeek-VL-7B-Chat	**50.00**	47.90	41.55	52.27	56.94	**25.00**	51.83	38.25	**99.33**	**83.43**
Qwen2-VL-7B	**54.28**	64.35	37.38	22.73	30.73	**34.00**	**84.78**	21.75	**54.32**	**72.00**
Qwen-VL-Chat	**52.33**	42.24	57.94	22.72	56.79	18.25	58.69	**70.87**	40.33	39.60
MoE-LLAVA-Phi2-2.7B-4e-384	**51.67**	35.34	50.47	31.82	43.88	**25.00**	59.78	38.84	44.17	31.20
mPLUG-Owl2-LLaMA2-7b	7.66	33.62	31.78	9.09	32.52	**27.21**	64.13	33.63	38.90	16.60
Phi-3.5-Vision-Instruct	**50.00**	55.17	**70.09**	31.81	**81.29**	24.00	63.04	53.33	0.00	45.80
Cambrian-1-8B	49.70	34.48	38.32	13.64	24.72	**25.00**	50.00	42.30	0.00	39.41
MiniGPT4-LLaMA2-7B	**79.87**	43.48	37.74	33.33	57.05	0.00	48.35	46.53	**93.96**	43.06
InternVL-Chat-V1-5	**54.30**	59.40	61.60	68.10	**79.40**	23.00	**76.00**	**84.20**	35.60	54.20
Mini-InternVL-Chat-4B-V1-5	42.00	40.51	50.46	22.70	**74.50**	21.20	57.60	36.84	15.33	43.60
InternLM-XComposer2-VL-1.8B	**51.00**	33.60	34.50	54.50	**78.30**	9.30	58.60	**76.30**	9.30	52.40
GPT4RoI	47.60	47.40	50.40	68.10	44.50	**31.00**	28.20	64.40	**62.30**	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	**54.33**	52.59	**67.29**	63.64	62.36	**28.75**	51.09	54.67	9.33	45.00
LLaVA-NeXT-34B	48.67	**65.52**	**73.83**	72.73	70.82	**32.50**	61.96	62.33	12.67	50.60
Pixtral 12B	**58.67**	45.69	63.55	36.37	67.48	**30.25**	57.61	59.67	8.33	39.40
SEED-LLaMA-13B	30.33	41.38	42.06	45.45	57.46	18.75	28.26	52.33	6.67	27.40
BLIP2	33.67	28.45	43.93	40.91	52.78	19.00	27.17	52.67	6.33	47.20
MiniMonkey	**50.00**	56.03	61.68	59.09	**77.28**	**29.75**	**75.00**	**76.32**	**65.00**	69.20
DeepSeek-VL-7B	48.33	46.55	38.32	50.00	56.79	**24.50**	52.17	37.33	**57.82**	**75.60**
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	**55.67**	57.76	66.36	**77.27**	70.60	**33.50**	55.43	62.00	22.33	52.20
ShareGPT4V-7B	47.67	42.24	52.34	59.09	57.24	**28.50**	46.74	53.67	15.33	44.80
ShareGPT4V-13B	45.67	49.14	57.94	72.72	61.47	**32.00**	51.09	55.67	17.33	46.20
BLIP-3 (XGen-MM)	46.00	54.31	58.88	63.64	67.04	**31.75**	57.61	61.33	19.00	50.60
AnyGPT	36.33	12.07	13.08	18.18	27.39	6.00	14.13	21.33	0.00	20.40
MiniCPM3-4B	**52.33**	56.03	57.94	63.64	**73.72**	**33.75**	66.30	64.00	9.67	54.00
LaVIT-V2 (7B)	36.60	35.34	36.45	40.91	57.91	19.00	26.09	49.00	7.67	26.80
GLM-VL-Chat	**55.33**	**66.38**	**71.96**	**77.27**	66.59	**30.50**	59.78	59.67	9.00	56.20
Gemini-1.5-Pro	**60.33**	**65.52**	**76.64**	**77.27**	70.16	**40.00**	61.96	67.00	41.00	70.40
Gemini-1.5-Flash	**58.33**	63.79	**73.83**	**77.27**	68.82	**33.75**	56.52	63.33	33.67	66.80
OMG-LLaVA-InternLM20B	1.67	3.45	4.67	9.09	2.45	2.00	3.26	2.33	2.67	3.20
Idefics3-8B-Llama3	**57.00**	46.55	**70.09**	63.64	71.27	**36.00**	64.13	63.33	32.00	61.20
NExT-GPT-V1.5	38.50	28.40	34.10	36.80	64.27	20.80	31.57	40.37	3.58	0.00
Vitron-V1	41.60	32.70	40.70	34.60	63.45	**26.50**	34.86	45.21	5.45	0.00
Otter	0.00	6.78	25.10	29.20	33.10	11.50	18.21	37.50	0.00	0.00
Show-o	**50.00**	1.72	1.87	0.00	2.00	0.25	28.26	62.99	50.00	1.20
NExT-Chat	39.81	44.79	35.29	50.00	38.46	21.69	59.78	43.64	11.48	37.45
Yi-vision-v2	31.33	**68.10**	**71.96**	22.73	**89.53**	**28.75**	35.87	61.54	45.33	50.23
Qwen2-VL-72B	40.43	1.72	52.55	60.95	58.22	18.42	68.93	69.58	**51.69**	70.37

Table 28: Results on Image Comprehension Group,#I-C-17, part C.

Model	#I-C-17 (Img VQA)									
	#21 ↑	#22 ↑	#23 ↑	#24 ↑	#25 ↑	#26 ↑	#27 ↑	#28 ↑	#29 ↑	#30 ↑
SoTA Specialist	70.60	59.30	61.90	61.40	62.80	62.10	70.20	83.80	82.40	68.60
GPT4V	68.60	22.81	20.40	24.10	22.64	22.02	18.01	**93.00**	**91.40**	**71.20**
GPT4o	**74.00**	37.30	40.30	43.45	35.47	39.87	29.60	**93.80**	**94.30**	65.00
GPT4o-mini	65.60	26.34	23.70	29.07	26.52	25.01	25.20	**90.50**	**89.50**	68.50
GPT-4o-4096	**77.00**	27.95	49.51	52.05	43.15	48.03	43.31	**94.40**	**95.60**	**69.60**
ChatGPT-4o-latest	**76.00**	28.61	40.01	42.66	35.43	37.96	36.81	**89.40**	**92.20**	**70.60**
Claude-3.5-Sonnet	59.95	39.10	45.29	48.88	44.27	54.41	59.85	76.11	**82.49**	**71.90**
Claude-3.5-Opus	58.08	42.24	45.67	47.91	42.24	53.51	62.86	76.00	**82.54**	**73.25**
Emu2-32B	44.80	24.20	30.80	24.60	22.40	36.20	44.60	71.40	72.60	60.40
DetGPT	36.80	21.40	16.90	20.20	19.50	17.40	17.90	50.60	54.80	54.40
InternVL2.5-8B	**100.00**	49.75	40.45	58.79	55.26	50.00	57.43	**97.10**	77.46	59.09
InternVL2.5-4B	**84.82**	33.70	35.99	54.69	53.03	33.92	57.38	79.13	79.84	64.82
InternVL2.5-2B	**83.54**	27.33	22.83	31.13	60.27	50.00	63.84	**94.11**	68.42	**68.90**
Monkey-10B-chat	**80.35**	39.10	42.29	**65.94**	57.99	42.91	47.34	**85.53**	80.15	44.84
DeepSeek-VL-7B-Chat	**84.58**	23.27	18.72	30.41	26.98	19.80	54.66	60.99	61.76	47.51
Qwen2-VL-7B	**73.60**	**75.37**	**71.96**	**82.41**	**78.84**	**69.40**	**77.60**	**92.10**	**94.74**	**79.20**
Qwen-VL-Chat	37.60	50.65	43.55	58.28	52.91	46.69	52.60	72.80	71.00	53.40
MoE-LLAVA-Phi2-2.7B-4e-384	30.60	**62.43**	58.87	**72.02**	**67.55**	57.35	32.63	42.80	43.00	41.80
mPLUG-Owl2-LLaMA2-7b	14.39	44.41	43.98	**65.46**	**63.62**	45.78	13.20	18.60	23.60	28.99
Phi-3.5-Vision-Instruct	44.20	**64.36**	**62.71**	**70.32**	**72.85**	60.24	**73.80**	**84.52**	**85.27**	57.20
Cambrian-1-8B	28.73	49.50	50.61	58.83	52.37	43.35	48.00	0.63	44.40	40.00
MiniGPT4-LLaMA2-7B	30.08	44.68	54.65	60.10	59.00	50.26	51.28	27.80	46.72	47.36
InternVL-Chat-V1-5	52.40	30.30	26.80	33.30	36.90	28.20	**73.40**	**87.70**	**83.30**	**70.60**
Mini-InternVL-Chat-4B-V1-5	43.00	29.80	26.20	40.10	37.60	24.80	62.00	80.20	65.60	63.00
InternLM-XComposer2-VL-1.8B	52.80	46.75	48.90	**69.80**	**65.40**	51.00	52.20	54.20	52.60	**74.00**
GPT4RoI	0.00	20.80	23.80	31.70	23.90	23.20	37.20	38.20	34.80	28.20
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	46.20	35.60	34.60	46.50	41.20	46.60	58.40	68.80	72.40	60.40
LLaVA-NeXT-34B	52.40	37.80	41.20	46.20	37.80	48.80	61.20	74.60	78.60	65.40
Pixtral 12B	53.40	33.20	38.20	47.90	32.40	51.40	48.60	64.20	74.20	66.20
SEED-LLaMA-13B	42.40	21.90	18.40	20.90	20.80	19.20	18.40	44.20	52.60	59.40
BLIP2	47.80	49.20	48.80	47.40	49.60	50.20	43.80	46.00	42.80	29.60
MiniMonkey	**71.60**	36.17	39.67	53.55	53.34	39.57	41.64	63.60	71.60	35.80
DeepSeek-VL-7B	**78.60**	21.81	17.29	30.06	22.76	17.44	53.60	61.80	60.80	43.00
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	56.60	35.30	40.75	47.60	43.57	49.72	58.40	65.20	69.40	64.80
ShareGPT4V-7B	45.20	27.53	32.93	37.34	35.26	40.95	49.80	57.80	61.00	55.40
ShareGPT4V-13B	48.20	29.23	35.92	41.29	38.29	42.94	52.00	60.40	63.60	57.80
BLIP-3 (XGen-MM)	52.00	32.42	38.32	45.82	41.87	46.80	54.60	62.80	65.20	62.40
AnyGPT	27.60	19.90	21.70	26.40	28.40	23.60	20.80	14.80	8.60	26.00
MiniCPM3-4B	57.80	36.70	38.90	50.70	48.70	48.60	63.20	62.80	67.60	62.80
LaVIT-V2 (7B)	43.20	29.40	24.50	33.70	36.70	31.90	35.20	42.00	46.80	48.40
GLM-VL-Chat	55.20	38.80	44.70	51.30	49.60	52.40	59.60	69.80	69.80	67.40
Gemini-1.5-Pro	64.80	35.90	34.50	41.80	39.80	36.50	**75.20**	**91.20**	**90.00**	63.60
Gemini-1.5-Flash	66.00	30.80	28.60	38.70	41.40	33.70	68.40	**89.60**	**86.80**	58.40
OMG-LLaVA-InternLM20B	2.60	21.10	21.30	22.20	17.50	12.40	3.20	5.40	4.20	2.80
Idefics3-8B-Llama3	59.80	29.40	36.70	38.40	36.50	32.80	62.60	**84.40**	**85.20**	62.20
NExT-GPT-V1.5	0.00	23.58	22.89	21.60	20.65	18.65	15.48	0.00	0.00	0.00
Vitron-V1	0.00	27.41	23.05	31.60	21.78	20.78	16.98	0.00	0.00	0.00
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.40	0.44	0.38	0.42	25.00	28.80	1.40	45.29	0.90	0.80
NExT-Chat	12.50	20.67	27.53	25.31	23.40	21.35	40.97	12.88	11.64	26.41
Yi-vision-v2	48.59	8.54	9.81	13.96	16.32	9.28	60.27	34.27	44.59	67.83
Qwen2-VL-72B	**74.51**	56.19	**68.02**	**86.06**	**81.24**	**71.33**	**78.17**	**90.91**	80.95	**83.57**

Table 29: Results on Image Comprehension Group,#I-C-17, part D.

Model	#I-C-17 (Img VQA)									
	#31 ↑	#32 ↑	#33 ↑	#34 ↑	#35 ↑	#36 ↑	#37 ↑	#38 ↑	#39 ↑	#40 ↑
SoTA Specialist	75.00	67.20	68.40	66.80	48.80	56.60	75.80	31.00	20.71	29.20
GPT4V	**77.10**	**70.40**	**73.90**	**82.80**	**89.40**	**98.20**	14.80	19.60	16.83	18.30
GPT4o	74.20	**71.80**	**72.10**	**83.70**	**92.40**	**91.80**	27.20	30.20	18.04	19.20
GPT4o-mini	70.50	35.80	39.00	**80.60**	**77.60**	**63.40**	10.60	7.80	15.23	18.70
GPT-4o-4096	75.00	**76.60**	**77.40**	**83.00**	**91.00**	**87.00**	25.80	**47.51**	15.83	0.00
ChatGPT-4o-latest	73.60	**74.30**	**76.60**	**82.40**	**92.00**	**63.00**	18.20	**34.60**	18.24	19.72
Claude-3.5-Sonnet	74.44	**67.53**	**71.63**	**67.13**	**86.20**	**84.12**	16.89	18.24	16.43	18.62
Claude-3.5-Opus	74.34	**68.98**	**70.97**	**66.84**	**82.41**	**84.20**	16.78	18.21	16.40	15.56
Emu2-32B	66.80	59.00	67.20	62.60	**71.40**	**74.60**	22.78	14.60	15.43	14.85
DetGPT	56.40	52.00	38.20	44.60	**54.60**	**60.60**	18.64	10.20	8.22	13.83
InternVL2.5-8B	36.11	52.05	50.74	33.71	15.54	18.54	12.40	6.90	2.21	3.17
InternVL2.5-4B	69.37	40.74	40.60	35.92	**49.00**	30.41	49.21	6.45	1.80	3.39
InternVL2.5-2B	68.93	49.99	46.86	64.26	3.18	1.74	8.25	5.74	3.55	4.59
Monkey-10B-chat	47.56	57.20	59.60	33.94	**50.40**	51.40	50.20	13.37	2.00	2.32
DeepSeek-VL-7B-Chat	50.05	51.09	56.04	28.06	3.50	6.50	11.57	7.93	1.93	4.11
Qwen2-VL-7B	**81.80**	13.33	41.91	**93.33**	**91.06**	**86.67**	45.80	**48.60**	**45.36**	20.89
Qwen-VL-Chat	54.60	3.20	22.00	61.00	47.00	50.40	45.80	15.40	17.89	5.07
MoE-LLAVA-Phi2-2.7B-4e-384	41.40	6.00	13.60	59.10	**51.20**	50.80	50.00	**37.00**	20.10	12.57
mPLUG-Owl2-LLaMA2-7b	25.60	30.20	5.20	14.80	**54.40**	51.60	44.40	**33.60**	0.00	9.11
Phi-3.5-Vision-Instruct	61.44	18.40	40.67	44.90	**48.97**	**62.46**	49.03	20.00	**32.06**	18.80
Cambrian-1-8B	64.17	9.70	11.00	49.30	43.98	**59.47**	43.70	29.20	4.30	16.52
MiniGPT4-LLaMA2-7B	55.68	9.80	18.70	43.70	0.67	48.32	**91.28**	**58.39**	18.92	4.13
InternVL-Chat-V1-5	69.20	55.00	55.20	**71.00**	**74.00**	**76.40**	60.20	27.60	0.00	3.00
Mini-InternVL-Chat-4B-V1-5	65.20	44.00	46.60	61.20	**61.40**	**65.00**	50.80	24.20	0.00	2.40
InternLM-XComposer2-VL-1.8B	73.40	64.60	62.60	**72.20**	**66.60**	49.00	48.80	20.00	1.60	5.80
GPT4RoI	31.60	10.60	8.20	28.00	**53.60**	50.40	50.00	11.80	0.00	14.60
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	64.20	58.60	59.60	52.40	**62.40**	**64.40**	30.20	12.20	12.02	16.42
LLaVA-NeXT-34B	70.60	62.80	65.60	60.40	**70.60**	**72.20**	42.40	16.40	15.63	17.93
Pixtral 12B	65.40	60.40	57.60	61.20	**68.20**	**68.20**	18.60	17.20	13.63	15.88
SEED-LLaMA-13B	50.20	44.40	35.80	40.80	**49.20**	**58.80**	11.40	11.60	8.62	11.73
BLIP2	31.20	11.40	10.80	40.40	**52.60**	**61.80**	59.40	6.80	1.40	2.10
MiniMonkey	37.80	46.80	57.40	49.60	48.20	53.60	61.80	28.33	0.00	3.41
DeepSeek-VL-7B	50.80	47.60	47.20	32.00	**51.60**	49.60	13.58	12.40	1.60	3.73
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	68.40	63.60	64.40	59.40	**69.20**	**67.80**	38.40	16.20	12.83	22.14
ShareGPT4V-7B	62.80	54.20	53.20	52.20	**63.20**	**59.60**	31.20	13.40	9.82	18.29
ShareGPT4V-13B	66.20	58.40	56.80	55.80	**65.80**	**62.60**	33.40	14.60	11.22	20.56
BLIP-3 (XGen-MM)	64.20	61.20	62.00	57.60	**67.40**	**64.40**	35.60	15.80	13.83	19.47
AnyGPT	42.20	14.60	6.40	24.80	0.00	34.00	16.40	3.40	5.81	0.00
MiniCPM3-4B	65.00	66.80	62.60	61.20	**64.20**	**72.60**	42.80	16.80	12.42	19.74
LaVIT-V2 (7B)	54.80	42.60	46.80	40.60	**50.80**	**57.20**	34.60	11.80	8.22	13.07
GLM-VL-Chat	70.80	65.20	63.40	63.80	**68.60**	**78.40**	44.00	17.40	15.83	26.34
Gemini-1.5-Pro	65.20	62.00	67.80	**74.40**	**82.60**	**89.60**	74.00	27.20	17.23	18.10
Gemini-1.5-Flash	61.00	53.60	65.60	**72.00**	**77.40**	**84.00**	68.80	20.80	14.63	16.30
OMG-LLaVA-InternLM20B	3.40	2.00	2.60	3.20	7.80	8.40	7.20	0.00	0.00	10.10
Idefics3-8B-Llama3	56.40	54.20	57.60	63.00	**74.40**	**76.20**	63.40	17.60	11.22	9.70
NExT-GPT-V1.5	0.00	20.60	0.00	28.60	26.50	36.90	51.40	12.60	2.70	14.60
Vitron-V1	0.00	11.40	0.00	39.40	27.80	**59.90**	60.70	9.40	1.80	13.47
Otter	0.00	0.00	0.00	0.00	**49.89**	0.00	0.00	0.00	0.00	0.00
Show-o	0.80	0.90	7.20	0.00	5.40	5.80	0.00	0.00	0.00	0.00
NExT-Chat	27.38	18.52	17.44	36.80	47.44	51.32	58.42	**49.85**	**47.33**	11.71
Yi-vision-v2	67.36	7.22	17.04	41.31	46.00	**65.80**	49.80	**48.60**	19.70	2.35
Qwen2-VL-72B	**88.93**	53.39	57.18	**87.76**	**96.32**	**91.25**	56.41	15.06	**48.21**	20.49

Table 30: Results on Image Comprehension Group,#I-C-18, part A.

Model	#I-C-18 (Ind-Anomaly Det)									
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑	#7 ↑	#8 ↑	#9 ↑	#10 ↑
SoTA Specialist	90.30	60.40	56.30	72.80	61.50	88.40	75.60	93.50	83.10	58.70
GPT4V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o-mini	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-4o-4096	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChatGPT-4o-latest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Sonnet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Opus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Emu2-32B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DetGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Monkey-10B-chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MoE-LLAVA-Phi2-2.7B-4e-384	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mPLUG-Owl2-LLaMA2-7b	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phi-3.5-Vision-Instruct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cambrian-1-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniGPT4-LLaMA2-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-Chat-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mini-InternVL-Chat-4B-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternLM-XComposer2-VL-1.8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-34B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pixtral 12B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SEED-LLaMA-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniMonkey	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP-3 (XGen-MM)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AnyGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniCPM3-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LaVIT-V2 (7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLM-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OMG-LLaVA-InternLM20B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Idefics3-8B-Llama3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	60.50	26.40	13.50	36.50	25.30	10.70	24.00	33.50	15.70	13.40
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Yi-vision-v2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 31: Results on Image Comprehension Group,#I-C-18, part B.

Model	#I-C-18 (Ind-Anomaly Det)								
	#11 ↑	#12 ↑	#13 ↑	#14 ↑	#15 ↑	#16 ↑	#17 ↑	#18 ↑	#19 ↑
SoTA Specialist	90.00	75.20	64.80	81.90	85.20	62.10	76.20	23.10	73.60
GPT4V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o-mini	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-4o-4096	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChatGPT-4o-latest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Sonnet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Opus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Emu2-32B	0.00	0.00	0.00	0.00	0.00	35.13	45.12	20.02	0.00
DetGPT	0.00	0.00	0.00	0.00	0.00	26.38	35.36	18.24	0.00
InternVL2.5-8B	0.00	0.00	0.00	0.00	0.00	37.00	25.00	**28.69**	62.50
InternVL2.5-4B	0.00	0.00	0.00	0.00	0.00	35.38	50.00	**40.44**	16.25
InternVL2.5-2B	0.00	0.00	0.00	0.00	0.00	34.13	20.71	**41.43**	45.00
Monkey-10B-chat	0.00	0.00	0.00	0.00	0.00	34.50	71.43	**40.03**	0.00
DeepSeek-VL-7B-Chat	0.00	0.00	0.00	0.00	0.00	36.75	28.57	**45.02**	41.43
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	41.12	60.71	**68.33**	35.00
Qwen-VL-Chat	0.00	0.00	0.00	0.00	0.00	26.75	71.42	**58.16**	5.00
MoE-LLAVA-Phi2-2.7B-4e-384	0.00	0.00	0.00	0.00	0.00	25.00	71.07	**42.63**	25.00
mPLUG-Owl2-LLaMA2-7b	0.00	0.00	0.00	0.00	0.00	28.49	56.79	**24.30**	68.75
Phi-3.5-Vision-Instruct	0.00	0.00	0.00	0.00	0.00	34.75	73.93	**60.96**	56.25
Cambrian-1-8B	0.00	0.00	0.00	0.00	0.00	22.75	60.71	10.96	37.50
MiniGPT4-LLaMA2-7B	0.00	0.00	0.00	0.00	0.00	15.63	70.00	**51.00**	57.50
InternVL-Chat-V1-5	0.00	0.00	0.00	0.00	0.00	**65.10**	**79.60**	**76.40**	0.00
Mini-InternVL-Chat-4B-V1-5	0.00	0.00	0.00	0.00	0.00	39.80	65.70	**69.30**	0.00
InternLM-XComposer2-VL-1.8B	0.00	0.00	0.00	0.00	0.00	36.50	71.40	**56.90**	0.00
GPT4RoI	0.00	0.00	0.00	0.00	0.00	**70.50**	70.30	**41.60**	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	0.00	0.00	0.00	0.00	0.00	41.00	39.29	**39.84**	0.00
LLaVA-NeXT-34B	0.00	0.00	0.00	0.00	0.00	38.89	49.29	**44.22**	0.00
Pixtral 12B	0.00	0.00	0.00	0.00	0.00	34.63	45.36	**46.22**	0.00
SEED-LLaMA-13B	0.00	0.00	0.00	0.00	0.00	26.13	37.14	**27.09**	0.00
BLIP2	0.00	0.00	0.00	0.00	0.00	17.23	47.14	**33.27**	0.00
MiniMonkey	0.00	0.00	0.00	0.00	0.00	26.41	7.86	**23.71**	0.00
DeepSeek-VL-7B	0.00	0.00	0.00	0.00	0.00	31.57	69.31	**47.26**	0.00
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	0.00	0.00	0.00	0.00	0.00	43.25	55.71	**48.41**	0.00
ShareGPT4V-7B	0.00	0.00	0.00	0.00	0.00	31.13	43.93	**39.44**	0.00
ShareGPT4V-13B	0.00	0.00	0.00	0.00	0.00	33.25	48.57	**41.24**	0.00
BLIP-3 (XGen-MM)	0.00	0.00	0.00	0.00	0.00	36.88	53.93	**41.04**	0.00
AnyGPT	0.00	0.00	0.00	0.00	0.00	17.75	37.50	17.53	0.00
MiniCPM3-4B	0.00	0.00	0.00	0.00	0.00	36.75	54.64	**37.25**	0.00
LaVIT-V2 (7B)	0.00	0.00	0.00	0.00	0.00	25.38	43.57	**25.30**	0.00
GLM-VL-Chat	0.00	0.00	0.00	0.00	0.00	40.25	57.50	**45.42**	0.00
Gemini-1.5-Pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OMG-LLaVA-InternLM20B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Idefics3-8B-Llama3	0.00	0.00	0.00	0.00	0.00	38.75	51.79	**40.04**	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	23.67	38.41	22.30	0.00
Vitron-V1	14.80	7.80	3.40	23.70	19.60	24.89	42.69	**26.10**	0.00
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	22.72	32.56	**39.25**	3.85
Yi-vision-v2	0.00	0.00	0.00	0.00	0.00	26.63	53.93	**53.98**	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	40.77	71.43	**49.46**	45.16

Table 32: Results on Image Comprehension Group, from #I-C-19 to #I-C-20.

Model	#I-C-19 (Med Seg)						#I-C-20 (Keypoint Det)
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑	#1 ↑
SoTA Specialist	50.10	16.30	25.85	23.55	53.40	37.88	95.70
GPT4V	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o-mini	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-4o-4096	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChatGPT-4o-latest	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Sonnet	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Opus	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Emu2-32B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DetGPT	0.00	0.00	0.00	0.00	0.00	35.26	0.00
InternVL2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Monkey-10B-chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MoE-LLAVA-Phi2-2.7B-4e-384	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mPLUG-Owl2-LLaMA2-7b	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phi-3.5-Vision-Instruct	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cambrian-1-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniGPT4-LLaMA2-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-Chat-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mini-InternVL-Chat-4B-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternLM-XComposer2-VL-1.8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	0.00	**17.70**	0.00	0.00	2.10	**41.50**	0.00
LLaVA-NeXT-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-34B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pixtral 12B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SEED-LLaMA-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniMonkey	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LISA	0.00	0.00	0.00	0.00	0.00	30.22	0.00
CogVLM-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP-3 (XGen-MM)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AnyGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniCPM3-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LaVIT-V2 (7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLM-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OMG-LLaVA-InternLM20B	1.68	14.53	19.96	17.75	1.02	0.00	0.00
Idefics3-8B-Llama3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	2.00	10.70	**35.60**	10.40	12.30	16.78	0.00
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Yi-vision-v2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 33: Results on Image Comprehension Group, from #I-C-21 to #I-C-22.

Model	#I-C-21 (Multi-img VQA)						#I-C-22 (MM Dialog)					
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑
SoTA Specialist	88.20	87.60	75.00	49.60	59.80	59.80	51.90	51.60	53.40	53.10	51.60	40.80
GPT4V	46.20	47.20	**82.40**	**94.00**	**84.80**	**76.80**	23.45	**55.06**	28.00	52.38	37.05	26.80
GPT4o	61.60	59.80	**95.60**	**98.20**	**85.00**	**87.60**	21.84	**62.09**	28.13	**53.13**	38.09	34.40
GPT4o-mini	26.80	28.40	**96.60**	**97.80**	**80.80**	**66.20**	22.79	47.74	24.94	42.82	32.13	11.60
GPT-4o-4096	51.40	50.00	**96.00**	**98.80**	**91.00**	**78.20**	20.88	**59.21**	30.22	50.20	40.12	2.40
ChatGPT-4o-latest	25.60	27.00	**95.80**	**97.60**	**76.20**	**79.20**	22.11	**54.44**	28.35	49.62	38.09	17.25
Claude-3.5-Sonnet	44.28	45.04	**90.56**	**96.56**	**83.38**	**75.94**	30.71	49.91	35.13	45.96	39.53	38.68
Claude-3.5-Opus	41.96	40.65	**89.46**	**94.03**	**81.52**	**74.70**	33.72	49.87	35.39	44.50	42.15	**41.51**
Emu2-32B	48.40	53.60	62.60	**66.80**	52.40	54.20	28.13	35.60	27.87	27.40	23.40	34.20
DetGPT	39.00	31.00	42.60	29.20	28.40	38.40	22.96	25.60	19.67	19.70	20.70	22.40
InternVL2.5-8B	50.00	49.43	23.91	21.26	18.94	17.21	17.52	30.35	27.96	35.52	29.30	29.54
InternVL2.5-4B	69.13	68.27	49.40	**50.84**	47.48	50.00	15.29	19.40	20.64	17.25	17.52	27.47
InternVL2.5-2B	46.79	43.02	7.18	7.30	2.62	53.00	17.31	28.54	28.46	18.99	20.63	28.09
Monkey-10B-chat	43.80	39.73	49.40	47.60	50.80	50.00	10.09	11.47	12.55	17.38	10.93	37.07
DeepSeek-VL-7B-Chat	59.73	56.20	8.92	7.86	8.58	4.02	14.36	17.99	33.17	19.04	28.30	27.59
Qwen2-VL-7B	78.83	77.52	**87.60**	**96.39**	**61.80**	**83.60**	20.50	47.01	33.01	**53.50**	42.38	31.20
Qwen-VL-Chat	35.60	33.60	64.80	**73.00**	**64.80**	45.80	18.26	30.91	25.80	39.69	27.22	21.09
MoE-LLAVA-Phi2-2.7B-4e-384	55.60	53.20	49.60	49.40	50.40	48.60	17.05	21.81	23.38	23.71	22.26	17.32
mPLUG-Owl2-LLaMA2-7b	55.40	52.40	53.00	**51.80**	47.20	49.80	18.38	30.96	25.06	31.02	27.30	8.90
Phi-3.5-Vision-Instruct	67.26	58.51	70.34	**88.00**	**62.80**	58.45	21.42	31.49	29.88	42.70	39.67	28.73
Cambrian-1-8B	48.44	50.31	59.12	**64.07**	45.46	23.97	8.53	26.70	26.29	18.90	9.37	22.80
MiniGPT4-LLaMA2-7B	**95.97**	85.91	12.08	8.05	45.64	1.34	18.96	18.93	26.05	23.41	18.97	17.49
InternVL-Chat-V1-5	63.40	84.20	53.00	**65.20**	55.60	45.60	11.60	17.20	17.20	19.10	16.80	**42.60**
Mini-InternVL-Chat-4B-V1-5	80.20	82.40	52.40	**59.60**	51.60	42.80	11.90	16.80	17.80	17.20	15.30	33.20
InternLM-XComposer2-VL-1.8B	75.20	68.80	50.00	49.20	46.90	38.50	20.50	23.40	22.00	26.10	24.10	34.70
GPT4RoI	54.60	53.60	8.20	25.80	50.00	35.20	15.10	28.00	27.40	39.90	24.50	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	54.60	34.20	52.60	**53.40**	41.40	57.60	22.16	38.40	23.04	37.40	37.40	29.80
LLaVA-NeXT-34B	62.40	51.20	66.40	**65.60**	58.20	**69.80**	27.24	42.40	25.18	40.20	36.40	32.20
Pixtral 12B	52.40	46.40	56.20	**57.80**	50.20	36.40	24.14	43.80	27.24	41.80	34.20	31.00
SEED-LLaMA-13B	41.80	28.40	44.40	28.60	25.60	**62.40**	17.25	34.40	16.85	24.20	20.80	21.60
BLIP2	57.20	58.40	48.60	44.40	53.00	58.80	35.40	36.00	32.80	33.60	35.00	39.60
MiniMonkey	69.00	43.40	59.40	**63.20**	33.60	**67.20**	15.63	13.52	16.89	15.26	10.19	33.20
DeepSeek-VL-7B	56.40	58.80	47.20	39.80	57.60	**61.40**	11.86	14.69	29.66	18.97	17.21	23.40
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	53.40	47.20	64.60	**62.40**	56.80	**67.20**	21.03	41.90	23.47	41.60	39.36	30.40
ShareGPT4V-7B	45.80	41.60	56.20	**56.20**	49.80	**61.60**	16.22	34.19	18.39	34.27	35.84	22.00
ShareGPT4V-13B	48.60	43.20	58.40	**57.40**	52.60	**62.40**	17.97	36.58	19.45	36.89	38.50	24.60
BLIP-3 (XGen-MM)	51.20	45.80	62.20	**60.80**	54.20	**65.80**	19.58	38.67	21.48	39.23	40.56	26.80
AnyGPT	0.00	0.00	27.60	12.40	6.80	14.20	11.34	18.40	8.64	17.10	17.40	6.80
MiniCPM3-4B	49.40	34.60	**78.00**	**64.00**	56.60	57.40	20.62	45.90	24.07	47.30	44.60	29.60
LaVIT-V2 (7B)	43.20	0.00	45.80	26.80	21.40	27.40	15.53	31.70	15.31	25.80	21.50	19.40
GLM-VL-Chat	54.20	50.40	67.40	**56.80**	42.80	**64.60**	22.47	42.40	27.39	46.40	43.10	31.40
Gemini-1.5-Pro	**88.80**	86.20	**82.20**	**84.60**	**82.60**	**83.00**	25.60	33.40	30.30	32.60	28.40	39.00
Gemini-1.5-Flash	84.20	81.40	**77.60**	**83.00**	**78.20**	**79.40**	22.80	30.30	27.40	34.00	25.70	33.60
OMG-LLaVA-InternLM20B	6.20	8.80	7.60	8.40	8.00	7.00	16.60	15.90	14.60	19.10	13.90	1.80
Idefics3-8B-Llama3	68.60	61.40	71.20	**59.80**	**64.80**	**63.40**	22.50	26.10	24.60	32.80	22.70	36.20
NExT-GPT-V1.5	57.60	57.30	47.90	45.80	51.60	16.40	18.60	28.40	28.70	31.76	22.36	36.81
Vitron-V1	59.20	60.30	48.60	47.40	55.80	28.70	18.70	29.60	25.60	35.29	18.65	**41.00**
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	50.60	49.80	48.20	43.40	46.80	49.60	0.20	0.10	0.10	0.10	0.10	0.00
NExT-Chat	33.79	28.95	52.40	**64.29**	**61.55**	13.79	16.32	17.81	19.15	20.38	20.05	21.37
Yi-vision-v2	50.40	49.80	61.80	**69.40**	**66.40**	**66.80**	12.12	18.30	25.74	20.58	25.82	7.75
Qwen2-VL-72B	80.41	80.15	**89.60**	**98.43**	**75.21**	**67.34**	24.95	48.33	33.34	50.70	44.23	30.01

Table 34: Results on Image Comprehension Group, #I-C-23.

Model	#I-C-23 (MM Reason)													
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑	#7 ↑	#8 ↑	#9 ↑	#10 ↑	#11 ↑	#12 ↑	#13 ↑	#14 ↑
SoTA Specialist	68.80	69.80	77.20	50.00	49.80	64.60	80.20	70.60	80.60	47.20	66.20	67.60	74.20	56.80
GPT4V	0.00	0.00	72.40	22.80	17.80	**91.00**	**91.20**	**79.00**	**81.80**	23.80	56.00	56.60	**80.90**	30.90
GPT4o	0.00	0.00	62.20	10.40	8.80	**97.40**	**97.40**	**75.80**	77.40	37.00	50.40	51.40	**81.40**	31.20
GPT4o-mini	0.00	0.00	76.40	18.80	19.20	**92.40**	**92.60**	**71.60**	66.80	17.80	56.40	53.20	73.20	34.10
GPT-4o-4096	0.00	40.60	75.00	7.80	6.60	**96.00**	**96.60**	**77.80**	79.20	31.60	51.00	52.00	**82.20**	39.00
ChatGPT-4o-latest	0.00	0.00	69.72	13.00	12.80	**93.28**	**93.80**	**76.20**	73.00	25.60	53.00	54.00	**82.00**	37.80
Claude-3.5-Sonnet	53.91	53.14	65.54	17.24	25.84	40.84	**81.56**	**74.82**	78.57	28.13	65.19	**67.70**	73.57	36.80
Claude-3.5-Opus	55.67	53.32	68.41	18.09	23.93	40.69	**84.83**	**72.19**	80.09	27.42	**67.96**	64.57	71.38	37.88
Emu2-32B	40.20	36.40	54.60	10.40	18.80	29.60	76.40	66.40	71.20	21.40	58.40	59.60	62.40	30.60
DetGPT	36.40	35.80	33.20	3.20	11.60	5.60	62.40	49.60	46.80	14.20	48.80	40.20	40.00	20.60
InternVL2.5-8B	33.33	45.06	**78.66**	34.18	34.83	34.18	62.71	58.45	57.98	**47.62**	37.98	37.84	42.41	39.46
InternVL2.5-4B	44.61	44.28	65.00	33.06	31.97	28.81	46.20	52.54	51.95	31.81	43.42	44.95	45.29	28.93
InternVL2.5-2B	11.63	45.45	60.20	33.50	33.68	32.72	70.83	65.98	43.25	34.02	36.14	29.62	46.42	40.53
Monkey-10B-chat	45.40	32.20	45.60	32.94	33.04	32.31	45.00	55.60	36.42	33.79	52.99	52.60	41.76	37.00
DeepSeek-VL-7B-Chat	46.52	68.60	56.20	39.84	39.64	41.70	55.40	31.82	39.97	37.03	38.72	39.45	42.71	56.60
Qwen2-VL-7B	52.00	51.40	**81.41**	43.40	44.00	43.40	57.71	**72.40**	42.20	**53.20**	59.35	59.49	68.20	13.33
Qwen-VL-Chat	37.60	40.00	62.20	49.20	49.80	31.20	45.80	45.50	14.20	**52.80**	20.00	20.70	58.19	40.20
MoE-LLAVA-Phi2-2.7B-4e-384	4.80	4.40	55.91	**54.00**	**52.60**	45.60	1.00	22.60	9.40	34.40	22.00	20.00	47.40	3.80
mPLUG-Owl2-LLaMA2-7b	53.00	62.40	55.71	28.00	27.80	28.20	55.20	56.30	29.20	14.60	39.40	41.20	35.00	**57.20**
Phi-3.5-Vision-Instruct	59.72	66.40	67.40	**51.06**	48.00	40.76	54.20	37.40	38.50	**54.00**	55.20	48.77	60.05	39.65
Cambrian-1-8B	36.54	49.62	56.20	44.36	39.81	29.74	46.00	39.10	19.82	46.60	40.27	19.20	37.60	25.78
MiniGPT4-LLaMA2-7B	39.83	44.34	48.90	39.46	35.90	33.26	39.83	38.69	20.00	**48.96**	41.28	22.17	27.61	28.00
InternVL-Chat-V1-5	46.40	47.60	61.80	**68.40**	**75.60**	**68.20**	56.40	61.60	61.00	**74.80**	53.00	53.20	68.20	43.00
Mini-InternVL-Chat-4B-V1-5	44.20	46.40	57.60	**53.40**	**69.80**	63.90	51.20	61.20	33.40	**55.20**	47.20	46.80	69.20	46.00
InternLM-XComposer2-VL-1.8B	48.20	45.40	60.60	41.90	**56.90**	53.70	45.40	53.20	48.60	**73.40**	54.00	53.00	54.80	37.00
GPT4RoI	39.20	22.60	38.60	0.00	0.00	0.00	27.20	50.20	19.60	0.00	47.00	47.40	28.40	36.40
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	49.20	47.00	54.60	7.20	14.40	28.80	68.40	63.40	68.20	16.60	54.20	49.20	60.60	26.80
LLaVA-NeXT-34B	51.00	53.40	60.20	10.80	17.40	37.60	75.60	68.40	70.40	20.20	56.80	56.20	65.20	28.60
Pixtral 12B	42.80	43.20	56.00	8.60	15.60	35.40	74.80	67.80	74.20	17.80	53.40	51.20	66.40	30.20
SEED-LLaMA-13B	43.20	39.80	46.80	2.80	4.20	8.40	64.80	37.80	42.40	10.20	48.40	44.60	46.60	17.40
BLIP2	43.00	44.40	53.40	1.60	2.00	0.20	41.60	58.80	30.20	0.00	46.20	43.40	34.80	37.80
MiniMonkey	42.40	29.80	42.88	29.80	18.80	29.40	40.40	52.40	31.20	28.00	48.60	44.20	36.40	37.00
DeepSeek-VL-7B	42.20	63.80	51.80	33.60	37.60	37.60	55.20	29.60	34.80	36.80	38.80	40.60	41.80	53.40
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	50.40	54.20	62.80	10.40	15.80	31.60	77.40	66.40	76.00	14.20	59.40	57.80	65.80	29.40
ShareGPT4V-7B	42.00	46.60	54.40	8.20	13.60	26.80	69.40	60.60	64.20	11.40	51.20	48.80	61.60	24.80
ShareGPT4V-13B	44.20	48.40	56.80	9.60	15.20	28.20	71.80	62.40	68.60	12.80	54.80	51.40	58.20	26.40
BLIP-3 (XGen-MM)	48.60	52.80	60.20	8.40	14.00	29.40	74.60	64.80	72.20	13.60	57.20	54.40	63.20	27.20
AnyGPT	26.80	31.20	24.80	1.40	0.00	0.00	32.40	2.80	26.40	1.20	10.60	12.40	21.40	8.60
MiniCPM3-4B	52.20	53.80	66.40	10.60	18.60	27.60	76.20	69.40	74.60	11.60	62.80	60.80	61.40	29.40
LaVIT-V2 (7B)	43.00	38.80	49.60	3.40	3.60	7.60	66.60	34.00	40.60	9.60	48.40	51.00	48.20	14.20
GLM-VL-Chat	50.60	57.40	64.80	12.20	14.60	32.40	79.80	70.20	79.80	15.40	61.00	65.20	66.80	28.80
Gemini-1.5-Pro	46.20	40.60	**77.40**	37.80	37.20	**91.60**	**93.20**	**76.40**	74.80	40.60	56.20	55.40	72.60	52.20
Gemini-1.5-Flash	44.80	38.20	69.20	33.00	33.60	**89.20**	**90.40**	**71.80**	73.00	37.80	48.40	52.20	68.40	46.80
OMG-LLaVA-InternLM20B	2.60	2.20	3.80	1.80	2.20	2.80	4.00	2.20	3.00	2.20	2.60	2.40	2.40	2.60
Idefics3-8B-Llama3	43.80	41.20	58.40	23.20	28.80	45.60	78.80	64.20	72.20	38.40	43.20	48.80	48.20	43.20
NExT-GPT-V1.5	19.60	18.40	18.60	10.70	24.78	28.96	10.36	0.00	37.98	**51.30**	13.40	3.40	10.30	34.60
Vitron-V1	38.41	20.57	21.50	9.45	22.16	31.20	14.89	0.00	28.64	**52.38**	17.50	4.80	14.60	35.80
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	42.80	45.60	25.00	7.20	0.00	9.00	45.00	33.20	15.20	0.00	39.80	41.00	11.60	37.00
NExT-Chat	37.08	61.76	57.80	13.71	13.36	25.66	33.93	45.37	25.61	35.98	15.73	29.35	23.68	44.30
Yi-vision-v2	45.09	37.80	70.03	26.40	14.40	36.62	61.23	49.79	21.90	24.75	58.08	57.06	35.16	39.49
Qwen2-VL-72B	53.06	55.56	**90.86**	46.67	37.13	49.32	52.23	**81.24**	54.42	**57.17**	64.46	52.34	**79.93**	51.51

Table 35: Results on Image Comprehension Group, from #I-C-24 to #I-C-25.

Model	#I-C-24 (Object Count)									#I-C-25 (NMT)
	#1 ↓	#2 ↓	#3 ↓	#4 ↓	#5 ↓	#6 ↓	#7 ↓	#8 ↓	#9 ↓	#1 ↑
SoTA Specialist	5.95	9.00	26.84	8.13	16.41	17.91	28.16	10.00	27.00	78.00
GPT4V	8.41	**3.20**	31.29	9.79	26.91	**10.72**	**10.86**	32.21	**11.14**	72.20
GPT4o	9.32	**2.30**	31.17	11.85	26.65	**10.76**	**5.29**	23.01	**15.70**	**93.00**
GPT4o-mini	10.06	**2.80**	30.98	13.62	43.87	**12.87**	**10.56**	26.30	**14.66**	**83.00**
GPT-4o-4096	7.47	23.66	31.11	14.65	21.75	**9.93**	**4.69**	16.26	**16.00**	**88.40**
ChatGPT-4o-latest	10.41	**2.60**	29.23	9.66	21.44	**10.41**	**5.86**	14.04	**22.35**	76.52
Claude-3.5-Sonnet	9.21	**1.89**	30.73	11.01	31.93	**11.08**	**8.24**	26.30	**13.42**	**82.23**
Claude-3.5-Opus	8.50	**0.24**	28.21	9.29	27.59	**7.29**	**8.40**	23.33	**11.99**	77.80
Emu2-32B	10.57	**1.85**	33.37	26.23	39.15	**14.24**	**17.58**	37.24	**12.83**	64.20
DetGPT	11.08	**1.54**	31.34	31.65	42.41	**13.58**	**16.28**	42.12	**12.94**	38.40
InternVL2.5-8B	9.11	**0.37**	35.40	10.84	27.17	**9.81**	**21.69**	23.68	**14.48**	7.82
InternVL2.5-4B	12.58	**0.59**	32.88	16.09	39.62	**15.38**	**27.88**	24.67	**14.41**	33.27
InternVL2.5-2B	14.11	**1.07**	37.27	37.15	69.25	20.03	**24.69**	28.67	**21.02**	20.00
Monkey-10B-chat	8.35	**1.04**	37.09	63.28	73.40	**13.91**	32.71	24.52	**14.68**	52.20
DeepSeek-VL-7B-Chat	25.84	**1.63**	36.38	69.02	100.00	24.81	34.30	24.52	**14.58**	5.34
Qwen2-VL-7B	12.49	**0.27**	30.30	41.30	46.49	**12.50**	**27.79**	16.00	**12.57**	**97.20**
Qwen-VL-Chat	7.41	**1.40**	**25.89**	36.60	67.19	**7.42**	**21.29**	16.45	**11.09**	**94.80**
MoE-LLAVA-Phi2-2.7B-4e-384	26.49	**1.05**	33.91	69.43	102.00	19.28	30.47	21.66	**14.33**	**80.80**
mPLUG-Owl2-LLaMA2-7b	14.28	**2.30**	35.22	55.12	69.30	**13.37**	**20.19**	21.57	**12.35**	**86.80**
Phi-3.5-Vision-Instruct	19.31	**0.80**	**26.80**	47.26	52.80	**10.12**	**22.80**	22.40	**11.03**	**90.10**
Cambrian-1-8B	14.08	**2.73**	39.15	50.61	95.19	**9.85**	**26.40**	22.06	**10.11**	71.40
MiniGPT4-LLaMA2-7B	9.60	**2.10**	41.20	65.30	**4.40**	20.03	28.49	20.58	**10.92**	76.32
InternVL-Chat-V1-5	14.30	**0.60**	34.60	41.60	69.60	**10.70**	**16.90**	23.90	**17.30**	**85.20**
Mini-InternVL-Chat-4B-V1-5	18.20	**1.94**	36.80	40.10	68.30	**14.10**	**22.10**	25.20	**17.25**	**80.40**
InternLM-XComposer2-VL-1.8B	13.30	**2.59**	39.00	54.80	62.30	**12.40**	**24.70**	26.50	**16.70**	**79.30**
GPT4RoI	∞	**3.29**	∞	∞	∞	∞	∞	∞	∞	0.00
GLaMM	∞	∞	∞	∞	∞	∞	∞	∞	∞	0.00
LLaVA-NeXT-13B	15.38	**3.81**	34.40	17.94	43.35	**13.38**	**22.44**	34.08	**15.24**	56.80
LLaVA-NeXT-34B	13.86	**3.60**	32.42	18.02	38.71	**14.46**	**20.06**	30.04	**14.53**	68.40
Pixtral 12B	12.07	**4.52**	34.65	18.52	45.04	**12.98**	**17.02**	36.14	**16.29**	60.40
SEED-LLaMA-13B	14.50	**5.40**	38.60	28.06	47.52	**15.02**	**16.42**	38.64	**20.49**	32.20
BLIP2	19.20	**2.67**	37.50	33.22	37.21	21.10	39.20	28.70	**20.60**	23.60
MiniMonkey	13.98	**3.78**	38.72	53.10	74.57	**15.80**	**23.64**	26.64	**23.62**	24.20
DeepSeek-VL-7B	27.29	**3.06**	38.61	62.59	68.83	22.63	33.49	21.59	**18.97**	5.80
LISA	∞	∞	∞	∞	∞	∞	∞	∞	∞	0.00
CogVLM-Chat	11.61	**3.02**	33.17	15.68	35.92	**11.63**	**18.34**	32.87	**19.35**	65.60
ShareGPT4V-7B	12.47	**4.15**	38.94	25.19	42.41	**13.19**	**22.68**	26.38	**21.94**	56.20
ShareGPT4V-13B	12.62	**2.84**	38.20	26.77	40.75	**12.22**	**21.54**	28.49	**23.58**	59.60
BLIP-3 (XGen-MM)	14.47	**3.54**	34.17	13.45	38.34	**14.21**	**23.58**	34.36	**17.62**	63.20
AnyGPT	13.92	**7.43**	40.82	42.57	62.53	**14.25**	30.94	36.07	**26.84**	14.60
MiniCPM3-4B	10.36	**3.24**	35.25	14.45	44.38	**10.30**	**19.64**	28.07	**15.03**	73.40
LaVIT-V2 (7B)	10.98	**4.62**	35.93	24.16	60.17	**10.41**	**15.64**	34.88	**18.21**	45.00
GLM-VL-Chat	9.09	**4.21**	33.44	17.28	34.36	**9.67**	**9.99**	27.88	**11.14**	68.80
Gemini-1.5-Pro	7.74	**0.32**	32.19	10.78	28.85	**9.90**	**8.28**	26.73	**12.03**	**86.40**
Gemini-1.5-Flash	11.23	**1.78**	33.03	12.36	33.17	**10.75**	**10.05**	25.00	**11.80**	**82.60**
OMG-LLaVA-InternLM20B	22.85	**2.89**	38.65	57.65	71.08	21.00	38.74	39.44	28.58	3.40
Idefics3-8B-Llama3	12.51	**3.71**	35.69	22.25	41.75	**12.25**	**13.68**	27.50	**16.84**	63.40
NExT-GPT-V1.5	11.58	**4.73**	36.77	31.25	67.40	**13.59**	**16.47**	30.58	**23.31**	0.00
Vitron-V1	12.34	**3.69**	35.91	27.59	56.70	**11.50**	**15.46**	28.15	**20.36**	0.00
Otter	22.04	**1.12**	38.34	60.17	91.58	27.75	36.11	27.67	**9.33**	0.00
Show-o	71.14	89.42	∞	∞	∞	72.13	∞	73.30	∞	0.00
NExT-Chat	29.56	10.03	36.84	69.78	89.27	21.55	**14.38**	28.68	**25.24**	43.52
Yi-vision-v2	13.63	**0.44**	29.83	24.51	44.48	**14.09**	**20.14**	13.23	**15.67**	8.77
Qwen2-VL-72B	9.64	**0.30**	27.13	9.81	42.55	**9.47**	**21.86**	10.40	**14.34**	**97.43**

Table 36: Results on Image Comprehension Group,#I-C-26, part A.

Model	#I-C-26 (Object Det)												
	#1 ↑	#2 ↑	#3 ↑	#4 ↑	#5 ↑	#6 ↑	#7 ↑	#8 ↑	#9 ↑	#10 ↑	#11 ↑	#12 ↑	#13 ↑
SoTA Specialist	87.20	87.80	32.60	41.10	73.80	29.00	34.00	62.40	93.20	26.20	37.60	18.10	61.60
GPT4V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o-mini	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-4o-4096	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChatGPT-4o-latest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Sonnet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Opus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Emu2-32B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DetGPT	0.00	85.99	30.69	39.93	67.60	25.20	26.38	56.40	88.80	3.00	33.40	9.00	55.60
InternVL2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Monkey-10B-chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MoE-LLAVA-Phi2-2.7B-4e-384	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mPLUG-Owl2-LLaMA2-7b	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phi-3.5-Vision-Instruct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cambrian-1-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniGPT4-LLaMA2-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-Chat-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mini-InternVL-Chat-4B-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternLM-XComposer2-VL-1.8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	37.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-34B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pixtral 12B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SEED-LLaMA-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniMonkey	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP-3 (XGen-MM)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AnyGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniCPM3-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LaVIT-V2 (7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLM-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OMG-LLaVA-InternLM20B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Idefics3-8B-Llama3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	12.34	37.84	**50.36**	**43.50**	**82.10**	16.40	**38.00**	56.80	75.80	**32.40**	31.70	**46.70**	60.34
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Yi-vision-v2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 37: Results on Image Comprehension Group,#I-C-26, part B.

Model	#I-C-26 (Object Det)												
	#14 ↑	#15 ↑	#16 ↑	#17 ↑	#18 ↑	#19 ↑	#20 ↑	#21 ↑	#22 ↑	#23 ↑	#24 ↑	#25 ↑	#26 ↑
SoTA Specialist	27.20	68.30	81.10	86.40	18.20	17.20	31.60	78.90	28.30	48.10	3.50	39.40	22.40
GPT4V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o-mini	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-4o-4096	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChatGPT-4o-latest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Sonnet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Opus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Emu2-32B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DetGPT	21.40	66.20	80.60	83.60	10.40	9.40	21.80	69.80	19.00	42.80	2.60	0.00	0.40
InternVL2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Monkey-10B-chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MoE-LLAVA-Phi2-2.7B-4e-384	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mPLUG-Owl2-LLaMA2-7b	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phi-3.5-Vision-Instruct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cambrian-1-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniGPT4-LLaMA2-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-Chat-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mini-InternVL-Chat-4B-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternLM-XComposer2-VL-1.8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-34B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pixtral 12B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SEED-LLaMA-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniMonkey	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP-3 (XGen-MM)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AnyGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniCPM3-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LaVIT-V2 (7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLM-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OMG-LLaVA-InternLM20B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Idefics3-8B-Llama3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	21.47	56.70	64.20	**87.36**	**23.40**	**36.50**	**42.70**	53.70	**34.60**	**56.10**	**10.40**	3.60	1.80
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Yi-vision-v2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 38: Results on Image Comprehension Group,#I-C-26, part C.

Model	#I-C-26 (Object Det)										
	#27 ↑	#28 ↑	#29 ↑	#30 ↑	#31 ↑	#32 ↑	#33 ↑	#34 ↑	#35 ↑	#36 ↑	#37 ↑
SoTA Specialist	67.60	63.40	10.90	48.60	38.00	97.10	61.10	97.50	25.90	58.20	64.20
GPT4V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4o-mini	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-4o-4096	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChatGPT-4o-latest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Sonnet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude-3.5-Opus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Emu2-32B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DetGPT	26.80	58.98	9.39	47.68	30.55	93.93	52.23	92.80	22.83	57.04	59.71
InternVL2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Monkey-10B-chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MoE-LLAVA-Phi2-2.7B-4e-384	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mPLUG-Owl2-LLaMA2-7b	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phi-3.5-Vision-Instruct	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cambrian-1-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniGPT4-LLaMA2-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-Chat-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mini-InternVL-Chat-4B-V1-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternLM-XComposer2-VL-1.8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-34B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pixtral 12B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SEED-LLaMA-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniMonkey	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LISA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ShareGPT4V-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLIP-3 (XGen-MM)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AnyGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniCPM3-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LaVIT-V2 (7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLM-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-1.5-Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OMG-LLaVA-InternLM20B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Idefics3-8B-Llama3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	37.50	**74.10**	**45.70**	**63.30**	26.40	85.70	56.40	87.60	**35.60**	47.80	26.70
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Yi-vision-v2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.24	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Image Generation Results. The complete results of all models on image generation are presented in Table 39 to Table 42.

Table 39: Results on Image Generation Group, from #I-G-1 to #I-G-3.

Model	#I-G-1 (Edge2Img Gen)	#I-G-2 (EEG2Img Gen)	#I-G-3 (Img Denose)						
	#1↓	#1↑	#1↑	#2↑	#3↑	#4↑	#5↑	#6↓	#7↑
SoTA Specialist	18.70	45.40	34.56	34.55	31.45	33.92	34.60	42.05	25.28
Emu2-32B	93.52	0.00	15.77	19.91	20.31	20.57	23.46	143.93	0.00
SEED-LLaMA-13B	127.10	0.00	15.29	16.85	18.45	18.72	19.71	170.65	0.00
AnyGPT	158.21	0.00	17.69	15.93	19.46	20.04	20.45	189.73	0.00
LaVIT-V2 (7B)	79.79	0.00	18.76	20.87	21.59	23.20	24.02	111.03	0.00
Next-GPT-V1.5	49.71	0.00	10.45	11.36	13.62	6.38	0.00	0.00	0.00
Vitron-V1	19.78	0.00	14.59	12.68	16.82	7.56	19.55	76.86	0.00
Show-o	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 40: Results on Image Generation Group, from #I-G-4 to #I-G-8.

Model	#I-G-4 (Img Enhance)						#I-G-5 (Img Inpaint)	#I-G-6 (Img-Style Trans)		#I-G-7 (Img2Mask Gen)	#I-G-8 (Img2Sketch Gen)
	#1↓	#2↑	#3↑	#4↑	#5↑	#6↑	#1↓	#1↓	#2↑	#1↑	#1↓
SoTA Specialist	3.29	27.35	24.49	0.54	22.16	19.97	4.86	16.73	31.26	99.29	15.06
Emu2-32B	∞	19.21	14.37	0.00	17.58	0.00	101.80	147.95	15.95	0.00	∞
SEED-LLaMA-13B	∞	15.77	13.17	0.00	15.09	0.00	127.42	179.87	16.79	0.00	∞
AnyGPT	∞	15.92	14.67	0.00	13.40	15.81	117.21	214.33	17.49	0.00	∞
LaVIT-V2 (7B)	∞	19.67	16.50	0.00	14.79	17.25	149.78	98.58	19.88	0.00	∞
Next-GPT-V1.5	20.45	2.63	0.00	0.00	0.00	0.00	75.71	65.89	16.51	0.00	47.30
Vitron-V1	15.89	3.78	17.86	0.15	6.90	0.00	32.15	51.48	19.17	86.53	23.47
Show-o	∞	0.00	0.00	0.00	0.00	0.00	∞	∞	0.00	0.00	∞

Table 41: Results on Image Generation Group, from #I-G-9 to #I-G-14.

Model	#I-G-9 (Img Edit)			#I-G-10 (Layout2Img Gen)	#I-G-11 (Mask2Img Gen)	#I-G-12 (Sketch2Img Gen)				#I-G-13 (Sound2Img Gen)	#I-G-14 (Text-based Img Edit)			
	#1↑	#2 ↑	#3↑	#1↓	#1↑	#1↓	#2↓	#3↓	#4↑	#1↑	#1↓	#2↑	#3↑	#4↓
SoTA Specialist	19.27	70.90	69.30	16.47	25.33	15.78	64.50	67.32	28.10	20.35	31.38	65.80	70.10	102.46
Emu2-32B	**22.72**	50.17	48.63	118.55	15.43	162.90	201.94	232.23	19.96	0.00	168.23	0.00	60.13	∞
SEED-LLaMA-13B	14.24	37.01	39.28	87.90	14.58	159.50	238.79	289.85	13.17	0.00	160.10	0.00	47.17	∞
AnyGPT	16.49	37.17	32.98	108.06	14.91	176.46	275.20	305.26	16.65	0.00	177.20	0.00	34.87	∞
LaVIT-V2 (7B)	17.81	60.78	60.61	89.78	15.79	123.75	246.68	256.77	18.96	0.00	144.22	0.00	56.51	∞
NExT-GPT-V1.5	2.60	45.38	36.59	86.45	6.53	32.67	89.39	75.86	15.76	12.45	76.59	38.65	60.69	∞
Vitron-V1	3.70	56.49	53.44	24.89	17.95	**13.76**	**36.45**	**57.59**	16.34	0.00	89.62	43.78	59.78	∞
Show-o	0.00	0.00	0.00	∞	0.00	∞	∞	∞	0.00	0.00	∞	0.00	0.00	∞

Table 42: Results on Image Generation Group, from #I-G-15.

Model	#I-G-15 (Txt2Img Gen)										
	#1↓	#2↑	#3↓	#4↓	#5↓	#6↓	#7↑	#8↑	#9↑	#10↓	#11↑
SoTA Specialist	19.86	32.57	10.58	15.66	177.30	9.71	31.20	27.86	29.91	12.52	30.07
Emu2-32B	87.59	22.97	71.44	∞	**167.91**	97.15	18.26	15.79	14.59	126.80	17.40
SEED-LLaMA-13B	78.91	26.53	84.25	∞	224.63	74.42	16.92	16.37	15.77	133.85	16.91
AnyGPT	94.82	17.64	88.13	∞	243.63	66.10	17.97	15.55	16.78	127.49	18.21
LaVIT-V2 (7B)	66.93	21.05	65.39	∞	181.58	57.33	21.22	16.48	17.44	101.92	19.12
NExT-GPT-V1.5	77.38	26.75	70.14	102.24	234.65	86.78	24.65	16.58	14.26	125.86	20.58
Vitron-V1	49.58	21.38	46.56	95.68	203.86	73.64	19.87	13.64	10.97	93.78	18.63
Show-o	120.36	9.21	119.19	220.46	290.44	199.99	10.44	7.55	8.2	181.22	9.73

C.7 Results of Video-related Tasks

Video Comprehension Results. The complete results of all models on video comprehension are presented from Table 43 to Table 55.

Table 43: Results on Video Comprehension Group, from #V-C-1 (a).

Model	#V-C-1 (Video QA)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
SoTA Specialist	27.70	9.80	17.80	20.50	24.80	25.70	26.40	25.20	21.20
InternVL-2.5-8B	21.10	7.50	13.10	12.40	16.20	24.50	21.70	18.50	14.90
InternVL-2-26B	**40.00**	**14.10**	**31.90**	**34.50**	20.30	16.10	22.60	18.60	**23.50**
InternVL-2.5-26B	25.60	8.20	15.10	19.50	22.30	**26.40**	25.10	23.70	17.80
Qwen2-VL-7B	27.70	7.40	16.40	19.80	22.30	25.10	25.20	16.80	15.70
Qwen2-VL-72B	27.70	9.80	17.80	**20.50**	**25.80**	**27.20**	26.10	24.50	18.90
DeepSeek-VL-2-small	9.70	7.30	2.60	6.10	10.40	7.60	8.40	10.70	9.30
DeepSeek-VL-2	8.50	6.50	5.00	5.80	11.30	10.50	8.20	10.10	8.60
LLaVA-One-Vision-7B	18.70	10.50	6.40	12.60	11.50	10.20	9.40	12.20	8.60
LLaVA-One-Vision-72B	18.50	9.40	3.10	12.30	17.10	15.80	15.10	17.60	16.60
Sa2VA-8B	24.40	10.30	**25.30**	21.20	12.60	13.40	18.40	14.80	19.60
Sa2VA-26B	25.30	**10.90**	**26.10**	**22.60**	15.20	14.90	19.60	17.80	20.10
CoLVA-2B	17.10	6.90	11.60	11.50	15.30	21.80	19.80	17.20	13.90
CoLVA-4B	19.60	7.30	13.00	12.20	16.10	24.40	21.50	18.20	14.80
Long-LLaVA-9B	**41.30**	**14.50**	**33.20**	**39.10**	16.50	13.60	19.60	18.60	18.60
NExT-GPT-V1.5	7.50	4.60	1.50	5.60	15.70	5.60	8.70	13.40	13.40
Vitron-V1	7.20	4.80	3.70	5.80	12.40	6.30	8.90	14.10	15.20
InternVL-2-8B	19.20	7.30	12.90	13.20	15.30	23.50	19.30	18.00	14.90
VidAgent	24.80	**10.80**	**25.90**	**22.20**	13.60	15.40	20.40	16.70	20.30

Table 44: Results on Video Comprehension Group, from #V-C-1 (b).

Model	#V-C-1 (Video QA)							
	#10↑	#11↑	#12↑	#13↑	#14↑	#15↑	#16↑	#17↑
SoTA Specialist	100.00	100.00	85.50	20.40	32.10	18.20	63.10	17.90
InternVL-2.5-8B	**100.00**	**100.00**	80.20	18.50	27.40	14.00	58.20	15.40
InternVL-2-26B	**100.00**	**100.00**	82.70	13.70	14.50	12.40	**79.30**	16.20
InternVL-2.5-26B	**100.00**	**100.00**	83.30	**22.50**	31.00	**19.20**	**69.20**	**20.60**
Qwen2-VL-7B	**100.00**	**100.00**	**85.50**	**20.40**	**32.10**	**18.20**	64.40	**18.30**
Qwen2-VL-72B	**100.00**	**100.00**	**86.90**	**22.30**	**32.30**	**19.40**	**70.20**	**20.40**
DeepSeek-VL-2-small	74.30	**100.00**	64.00	7.40	2.10	13.60	49.30	8.40
DeepSeek-VL-2	53.50	**100.00**	72.50	4.80	1.90	1.50	46.70	10.40
LLaVA-One-Vision-7B	**100.00**	**100.00**	64.60	4.50	11.60	6.80	**69.00**	10.10
LLaVA-One-Vision-72B	**100.00**	**100.00**	70.00	6.20	13.60	13.50	**83.70**	17.40
Sa2VA-8B	**100.00**	**100.00**	80.80	12.70	13.30	7.70	77.70	12.10
Sa2VA-26B	**100.00**	**100.00**	83.10	16.80	20.40	14.10	**79.60**	14.10
CoLVA-2B	**100.00**	**100.00**	76.00	17.20	16.60	8.60	54.20	12.60
CoLVA-4B	**100.00**	**100.00**	80.00	18.10	26.50	13.10	56.30	14.50
Long-LLaVA-9B	**100.00**	**100.00**	76.20	11.50	12.90	10.10	72.30	16.40
NExT-GPT-V1.5	83.50	**100.00**	64.50	3.20	8.90	1.20	51.70	4.30
Vitron-V1	**100.00**	98.00	73.80	3.70	7.10	3.50	56.70	5.20
InternVL-2-8B	100.00	99.00	78.6	17.50	27.90	14.70	58.70	15.70
VidAgent	**100.00**	**100.00**	82.80	14.20	24.30	10.70	**75.70**	12.30

Table 45: Results on Video Comprehension Group, from #V-C-2 (a).

Model	#V-C-2 (Vid-Obj Recog)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
SoTA Specialist	11.10	21.00	29.10	4.50	65.90	10.50	66.70	64.10	76.60
InternVL-2.5-8B	8.50	**21.50**	22.10	3.90	56.20	8.10	62.20	60.80	**77.90**
InternVL-2-26B	**17.10**	**24.50**	2.60	**8.40**	**88.90**	**15.00**	**78.90**	**75.00**	**82.20**
InternVL-2.5-26B	**13.20**	**25.10**	25.50	**4.60**	65.20	**13.20**	**75.30**	**70.20**	**80.20**
Qwen2-VL-7B	**13.10**	20.30	**29.10**	**4.50**	**65.90**	**10.50**	61.30	63.30	63.20
Qwen2-VL-72B	**15.80**	**24.80**	**32.10**	**4.50**	65.20	**12.20**	**73.20**	**74.50**	70.50
DeepSeek-VL-2-small	12.00	10.70	0.00	**4.80**	18.60	8.70	50.00	50.00	44.40
DeepSeek-VL-2	9.00	8.70	2.90	**6.50**	30.00	8.50	56.70	40.00	50.00
LLaVA-One-Vision-7B	**13.80**	**22.30**	11.50	**5.00**	46.50	5.00	51.10	53.00	52.20
LLaVA-One-Vision-72B	**16.70**	**24.20**	17.40	**8.80**	50.50	6.90	**80.00**	**80.00**	**77.80**
Sa2VA-8B	10.00	9.20	0.00	5.30	0.00	**12.70**	**80.00**	**76.00**	**80.00**
Sa2VA-26B	**12.80**	12.50	0.00	**6.40**	0.00	**13.20**	**82.40**	**78.50**	**80.20**
CoLVA-2B	7.20	17.70	17.40	3.90	50.20	6.50	56.00	54.40	71.60
CoLVA-4B	8.10	20.10	20.00	3.70	55.40	7.10	60.00	58.80	75.00
Long-LLaVA-9B	**16.60**	**29.30**	1.50	**9.40**	4.00	**14.00**	72.20	67.00	71.10
NExT-GPT-V1.5	4.60	6.10	0.00	1.20	12.10	5.30	64.80	53.10	68.40
Vitron-V1	3.80	7.20	0.00	2.40	14.00	6.10	70.10	68.30	**77.10**
InternVL-2-8B	7.80	20.50	20.80	4.20	55.80	7.10	60.70	60.60	75.60
VidAgent	**11.30**	13.30	14.20	**6.30**	53.20	6.50	63.20	**74.30**	**80.00**

Table 46: Results on Video Comprehension Group, from #V-C-2 (b).

Model	#V-C-2 (Vid-Obj Recog)							
	#10↑	#11↑	#12↑	#13↑	#14↑	#15↑	#16↑	#17↑
SoTA Specialist	60.10	57.70	23.00	24.70	92.80	91.00	71.30	73.80
InternVL-2.5-8B	55.30	50.70	18.60	22.40	0.00	0.00	0.00	0.00
InternVL-2-26B	**65.70**	**59.40**	22.00	24.60	0.00	0.00	0.00	0.00
InternVL-2.5-26B	**65.70**	**59.40**	22.00	24.60	0.00	0.00	0.00	0.00
Qwen2-VL-7B	58.10	52.10	17.90	21.30	0.00	0.00	0.00	0.00
Qwen2-VL-72B	**67.20**	**62.00**	22.90	24.50	0.00	0.00	0.00	0.00
DeepSeek-VL-2-small	36.70	46.70	5.40	8.90	0.00	0.00	0.00	0.00
DeepSeek-VL-2	42.20	42.20	10.50	14.90	0.00	0.00	0.00	0.00
LLaVA-One-Vision-7B	58.90	**66.70**	7.40	6.40	0.00	0.00	0.00	0.00
LLaVA-One-Vision-72B	**73.30**	**74.40**	12.50	9.80	0.00	0.00	0.00	0.00
Sa2VA-8B	**63.30**	**80.00**	6.00	4.40	0.00	0.00	0.00	0.00
Sa2VA-26B	**65.60**	**82.20**	8.00	5.80	0.00	0.00	0.00	0.00
CoLVA-2B	50.40	48.20	14.40	19.80	0.00	0.00	0.00	0.00
CoLVA-4B	53.30	50.10	17.60	20.50	0.00	0.00	0.00	0.00
InternVL-2-26B	**72.20**	**80.00**	3.70	4.80	0.00	0.00	0.00	0.00
Long-LLaVA-9B	**66.70**	65.60	15.00	13.80	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	57.20	64.70	10.70	9.60	0.00	0.00	0.00	0.00
Vitron-V1	**60.80**	65.20	11.90	11.30	0.00	0.00	0.00	0.00
InternVL-2-8B	55.60	50.90	19.60	21.40	0.00	0.00	0.00	0.00
VidAgent	**63.30**	**80.00**	6.50	14.40	0.00	0.00	0.00	0.00

Table 47: Results on Video Comprehension Group, from #V-C-3 (a).

Model	#V-C-3 (Vid Act Recog)							
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑
SoTA Specialist	17.60	14.60	20.10	17.40	21.30	13.50	17.70	26.00
InternVL-2.5-8B	13.20	10.50	14.20	15.10	17.40	6.80	14.90	22.60
InternVL-2-26B	**19.40**	13.70	19.80	15.80	20.80	12.40	17.30	24.10
InternVL-2.5-26B	**19.40**	13.70	19.80	15.80	20.80	12.40	17.30	24.10
Qwen2-VL-7B	14.10	9.80	13.90	14.80	18.40	9.00	15.20	13.50
Qwen2-VL-72B	**20.80**	12.90	**20.20**	16.90	**23.00**	**15.90**	**17.80**	18.90
DeepSeek-VL-2-small	7.40	4.30	7.90	14.10	8.50	6.30	14.30	13.70
DeepSeek-VL-2	14.60	7.90	14.00	**17.70**	10.80	**15.50**	**21.00**	19.60
LLaVA-One-Vision-7B	10.10	8.60	11.20	10.60	11.10	10.80	9.90	13.70
LLaVA-One-Vision-72B	**20.60**	**16.20**	**22.10**	17.10	**23.00**	**15.80**	17.40	21.50
Sa2VA-8B	**17.80**	**22.40**	**22.50**	**22.40**	**21.90**	**19.80**	**25.30**	**27.50**
Sa2VA-26B	**18.80**	**22.60**	**24.10**	**24.60**	**23.10**	**20.40**	**26.40**	**28.20**
CoLVA-2B	11.30	9.20	10.90	13.90	14.40	5.30	12.20	20.00
CoLVA-4B	12.40	9.90	13.00	14.30	16.60	6.10	13.80	21.10
InternVL-2-26B	**20.20**	**20.50**	**25.50**	**24.40**	20.60	**27.30**	**29.10**	23.30
Long-LLaVA-9B	**21.90**	**18.90**	**21.00**	**20.00**	18.70	**22.40**	**23.10**	21.60
NExT-GPT-V1.5	16.50	**14.90**	16.50	10.70	17.60	11.60	8.70	12.60
Vitron-V1	13.60	**16.20**	17.40	13.50	17.20	12.00	9.10	10.80
InternVL-2-8B	12.20	10.60	14.20	15.10	17.30	6.80	13.80	21.50
VidAgent	12.30	14.40	12.30	**22.30**	20.80	**19.80**	**22.30**	23.50

Table 48: Results on Video Comprehension Group, from #V-C-3 (b).

Model	#V-C-3 (Vid Act Recog)							
	#9↑	#10↑	#11↑	#12↑	#13↑	#14↑	#15↑	#16↑
SoTA Specialist	18.00	24.90	20.40	61.30	17.20	6.50	19.60	24.80
InternVL-2.5-8B	15.70	22.00	15.80	0.00	16.80	**14.20**	14.30	18.70
InternVL-2-26B	17.40	**25.40**	**21.20**	0.00	**24.30**	**18.90**	**23.40**	**25.40**
InternVL-2.5-26B	17.40	**25.40**	**21.20**	0.00	**24.30**	**18.90**	**23.40**	**25.40**
Qwen2-VL-7B	14.90	22.80	14.80	0.00	18.40	**13.60**	13.80	19.20
Qwen2-VL-72B	**19.20**	**26.20**	**22.30**	0.00	**25.90**	**19.80**	**21.10**	**28.70**
DeepSeek-VL-2-small	8.70	13.20	7.40	0.00	15.50	**14.20**	16.40	0.80
DeepSeek-VL-2	14.80	17.10	12.20	0.00	11.50	4.90	5.30	6.40
LLaVA-One-Vision-7B	10.90	12.70	10.90	0.10	15.00	3.20	12.70	8.00
LLaVA-One-Vision-72B	**18.40**	22.50	**20.70**	0.20	**26.90**	**12.50**	**19.60**	**30.40**
Sa2VA-8B	**23.20**	**26.50**	20.00	0.00	11.30	**7.40**	0.00	0.00
Sa2VA-26B	**23.60**	**26.80**	**20.80**	0.00	13.10	**8.80**	0.00	0.00
CoLVA-2B	12.80	18.90	13.20	0.00	14.20	**11.90**	13.10	15.30
CoLVA-4B	13.60	19.80	15.40	0.00	16.20	**13.70**	14.00	16.90
InternVL-2-26B	**21.00**	21.80	19.40	0.00	13.60	**7.60**	19.10	**37.60**
Long-LLaVA-9B	**21.10**	22.20	**21.00**	0.40	12.20	9.70	0.00	0.00
NExT-GPT-V1.5	14.20	6.30	9.20	0.00	8.50	4.10	0.00	0.00
Vitron-V1	16.50	7.40	10.50	0.00	7.80	4.70	0.00	0.00
InternVL-2-8B	14.90	22.30	16.80	0.00	15.80	14.80	14.00	17.70
VidAgent	**20.30**	24.50	**20.40**	0.00	12.30	**8.80**	13.00	15.30

Table 49: Results on Video Comprehension Group, from #V-C-4 to #V-C-5.

Model	#V-C-4 (Vid Understand)									#V-C-5 (IW VOS)			
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑	#1↑	#2↑	#3↑	#4↑
SoTA Specialist	23.80	21.50	26.70	23.00	25.90	14.50	16.70	25.80	29.60	78.30	82.90	86.60	79.60
InternVL-2.5-8B	22.60	18.70	21.20	19.30	22.90	8.40	14.20	19.80	22.40	0.00	0.00	0.00	0.00
InternVL-2.5-26B	**24.90**	19.60	24.20	22.10	24.10	10.30	15.80	24.20	28.90	0.00	0.00	0.00	0.00
Qwen2-VL-7B	19.70	17.90	24.80	20.10	21.90	9.70	13.90	16.80	25.80	0.00	0.00	0.00	0.00
Qwen2-VL-72B	22.20	**24.80**	25.10	21.80	**26.70**	11.70	14.00	25.30	**32.70**	0.00	0.00	0.00	0.00
DeepSeek-VL-2-small	15.30	5.00	9.10	6.40	9.10	4.30	5.40	10.50	5.80	0.00	0.00	0.00	0.00
DeepSeek-VL-2	16.20	6.70	13.40	9.20	13.20	9.20	14.40	10.10	16.60	0.00	0.00	0.00	0.00
LLaVA-One-Vision-7B	7.50	4.10	12.10	12.30	6.30	2.00	3.40	9.10	6.10	0.00	0.00	0.00	0.00
LLaVA-One-Vision-72B	8.70	5.80	24.40	19.30	8.30	3.80	5.80	11.60	7.40	0.00	0.00	0.00	0.00
Sa2VA-8B	4.30	4.60	**30.00**	14.90	5.60	4.00	4.30	5.30	5.00	0.00	0.00	0.00	0.00
Sa2VA-26B	6.60	6.90	**32.20**	16.90	6.20	5.60	5.80	6.90	6.40	0.00	0.00	0.00	0.00
CoLVA-2B	18.20	13.70	19.30	17.30	18.20	5.40	12.40	16.90	19.70	0.00	0.00	0.00	0.00
CoLVA-4B	20.10	17.70	21.00	18.70	20.40	7.90	12.90	18.70	21.20	0.00	0.00	0.00	0.00
Long-LLaVA-9B	13.90	14.70	14.00	16.80	14.70	**16.20**	**22.40**	14.30	12.80	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	1.40	5.40	13.40	7.80	0.00	3.40	5.00	14.50	3.50	0.00	0.00	0.00	0.00
Vitron-V1	2.30	5.70	14.10	6.20	0.00	3.70	5.70	13.60	3.90	59.80	60.20	54.20	36.40
InternVL-2-8B	22.30	17.60	20.20	19.30	20.30	8.40	14.20	18.80	17.40	0.00	0.00	0.00	0.00
VidAgent	20.30	14.60	**32.20**	15.40	13.60	4.30	14.30	15.30	12.30	0.00	0.00	0.00	0.00

Table 50: Results on Video Comprehension Group, from #V-C-6 to #V-C-9.

Model	#V-C-6 (General VOS)				#V-C-7 (Street VOS)			#V-C-8 (RVOS)			#V-C-9 (ReVOS)	
	#1↑	#2↑	#3↑	#4↑	#1↑	#2↑	#3↑	#1↑	#2↑	#3↑	#1↑	#2↑
SoTA Specialist	92.60	88.60	89.40	71.10	64.70	50.50	48.40	69.70	72.40	52.40	40.70	40.60
InternVL-2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2.5-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-2-small	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-One-Vision-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-One-Vision-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sa2VA-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**80.40**	**73.60**	**59.10**	**55.00**	**46.90**
Sa2VA-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CoLVA-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CoLVA-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Long-LLaVA-9B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	74.20	64.10	45.30	63.10	12.70	20.40	6.80	68.40	63.80	50.30	**45.20**	**41.30**
InternVL-2-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VidAgent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 51: Results on Video Comprehension Group, from #V-C-10 to #V-C-13.

Model	#V-C-10 (Temp Act Det)		#V-C-11 (C-ReVOS)					#V-C-12 (Vid-Ground)			#V-C-13 (Vid-Depth Est)			
	#1↑	#2↑	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#2↑	#3↑	#1↓	#2↓	#3↓	#4↓
SoTA Specialist	26.00	35.60	40.20	73.00	35.10	38.70	42.20	31.30	16.30	21.60	0.27	0.12	0.07	0.10
InternVL-2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
InternVL-2.5-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
DeepSeek-VL-2-small	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
DeepSeek-VL-2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
LLaVA-One-Vision-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
LLaVA-One-Vision-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
Sa2VA-8B	0.00	0.00	**58.40**	**86.20**	**41.20**	**58.90**	**56.70**	0.00	0.00	0.00	∞	∞	∞	∞
Sa2VA-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
CoLVA-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
CoLVA-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
InternVL-2-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
Long-LLaVA-9B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
Vitron-V1	6.70	3.60	**51.70**	**81.50**	33.60	36.90	**53.90**	**36.40**	15.70	10.80	∞	∞	∞	∞
InternVL-2-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞
VidAgent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞	∞

Table 52: Results on Video Comprehension Group, from #V-C-14.

Model	#V-C-14 (Obj Match)							
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑
SoTA Specialist	40.90	31.00	47.60	76.60	50.60	51.20	39.80	46.80
InternVL-2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2.5-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-2-small	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-One-Vision-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-One-Vision-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sa2VA-8B	13.60	20.70	22.40	24.70	11.60	24.00	23.30	18.50
Sa2VA-26B	24.80	14.30	33.60	43.80	16.80	31.20	31.80	31.00
CoLVA-2B	**40.90**	**31.00**	**47.70**	68.80	**50.60**	49.60	33.50	38.40
CoLVA-4B	38.30	**31.00**	41.10	**76.60**	41.70	**51.20**	**39.80**	**46.80**
InternVL-2-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Long-LLaVA-9B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VidAgent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 53: Results on Video Comprehension Group, from #V-C-15 to #V-C-16.

Model	#V-C-15 (Obj Track)							#V-C-16 (Long-Vid Track)				
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#1↑	#2↑	#3↑	#4↑	#5↑
SoTA Specialist	66.60	81.80	71.10	69.50	72.20	61.10	60.40	76.30	18.30	82.90	80.20	60.40
InternVL-2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2.5-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-2-small	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-One-Vision-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-One-Vision-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sa2VA-8B	22.60	77.80	56.70	49.30	38.20	14.30	5.90	61.20	4.60	71.80	69.20	25.00
Sa2VA-26B	24.10	78.30	59.10	49.60	38.10	15.60	7.60	63.30	5.10	71.60	69.60	25.90
CoLVA-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CoLVA-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Long-LLaVA-9B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VidAgent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 54: Results on Video Comprehension Group, from #V-C-17 to #V-C-18.

Model	#V-C-17 (UAV Track)					#V-C-18 (UW Track)				
	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#2↑	#3↑	#4↑	#5↑
SoTA Specialist	80.60	78.50	59.90	84.20	81.90	78.60	78.90	77.10	76.90	63.90
InternVL-2.5-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2.5-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-2-small	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-VL-2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-One-Vision-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-One-Vision-72B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sa2VA-8B	56.80	47.70	25.20	23.80	59.40	44.50	58.50	64.80	43.20	29.10
Sa2VA-26B	57.80	48.10	25.40	24.70	59.60	44.70	58.60	65.30	43.70	29.80
CoLVA-2B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CoLVA-4B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2-26B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Long-LLaVA-9B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-2-8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VidAgent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 55: Results on Video Comprehension Group, from #V-C-19 to #V-C-20.

Model	#V-C-19 (UAV Track)			#V-C-20 (UW Track)			
	#1↓	#2↓	#3↓	#1↑	#2↑	#3↑	#4↑
SoTA Specialist	49.20	39.40	23.00	22.70	71.80	48.51	32.98
InternVL-2.5-8B	∞	∞	∞	19.40	0.00	0.00	0.00
InternVL-2.5-26B	∞	∞	∞	21.20	0.00	0.00	0.00
Qwen2-VL-7B	∞	∞	∞	19.90	0.00	0.00	0.00
Qwen2-VL-72B	∞	∞	∞	22.80	0.00	0.00	0.00
DeepSeek-VL-2-small	∞	∞	∞	6.10	0.00	0.00	0.00
DeepSeek-VL-2	∞	∞	∞	12.90	0.00	0.00	0.00
LLaVA-One-Vision-7B	∞	∞	∞	5.10	0.00	0.00	0.00
LLaVA-One-Vision-72B	∞	∞	∞	6.80	0.00	0.00	0.00
Sa2VA-8B	∞	∞	∞	5.90	0.00	0.00	0.00
Sa2VA-26B	∞	∞	∞	6.80	0.00	0.00	0.00
CoLVA-2B	∞	∞	∞	14.40	0.00	0.00	0.00
CoLVA-4B	∞	∞	∞	16.90	0.00	0.00	0.00
InternVL-2-26B	∞	∞	∞	3.80	0.00	0.00	0.00
Long-LLaVA-9B	∞	∞	∞	16.80	0.00	0.00	0.00
NExT-GPT-V1.5	∞	∞	∞	0.20	0.00	0.00	0.00
Vitron-V1	∞	∞	∞	0.10	0.00	0.00	0.00
InternVL-2-8B	∞	∞	∞	19.4	0.00	0.00	0.00
VidAgent	∞	∞	∞	15.5	0.00	0.00	0.00

Video Generation Results. The complete results of all models on video generation are presented from Table 56 to Table 59.

Table 56: Results on Video Generation Group, from #V-G-1.

Model	#V-G-1 (Txt2Vid Gen)												
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑	#10↑	#11↑	#12↑	#13↓
SoTA Specialist	96.88	97.03	99.72	64.20	60.92	69.13	83.62	79.44	47.82	48.20	55.08	24.86	71.27
VidAgent	94.58	95.79	97.75	38.40	54.40	53.53	68.26	48.23	0.00	0.00	34.91	22.38	73.25
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞
NExT-GPT-V1.5	43.20	36.40	55.10	12.00	13.40	30.70	36.10	10.70	0.00	0.00	6.90	6.10	97.56
Vitron-V1	63.50	46.70	68.40	15.70	15.60	47.80	58.90	13.80	0.00	0.00	28.40	12.50	106.30

Table 57: Results on Video Generation Group, from #V-G-2 and #V-G-3.

Model	#V-G-2 (Condt Vid Gen)			#V-G-3 (Act Txt2Vid Gen)				
	#1↑	#2↑	#3↑	#1↑	#2↓	#3↓	#4↓	#5↓
SoTA Specialist	25.41	90.75	51.20	95.61	72.74	105.62	95.36	75.36
VidAgent	22.36	83.43	37.40	53.20	87.38	111.62	102.62	89.40
LM4LV	0.00	0.00	0.00	0.00	∞	∞	∞	∞
NExT-GPT-V1.5	3.45	2.40	14.30	2.60	126.74	140.70	254.12	126.94
Vitron-V1	6.67	33.80	17.50	44.90	98.50	135.90	186.47	115.78

Table 58: Results on Video Generation Group, from #V-G-4.

Model	#V-G-4 (Img2Vid Gen)										
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑	#10↑	#11↑
SoTA Specialist	59.46	61.55	60.79	62.89	67.37	63.17	59.21	61.97	62.18	65.45	67.89
VidAgent	**60.45**	**62.31**	**61.95**	**62.89**	67.19	62.94	**60.13**	**65.14**	**63.35**	64.19	**69.20**
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-GPT-V1.5	10.53	18.76	20.91	17.53	16.79	14.63	18.52	13.84	14.52	16.51	13.74
Vitron-V1	23.40	36.10	15.20	36.40	47.80	26.10	23.70	18.50	17.90	16.57	14.37

Table 59: Results on Video Generation Group, from #V-G-5 and #V-G-6.

Model	#V-G-5 (Vid Enhance)									#V-G-6 (Vid Edit)				
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑	#1↑	#2↑	#3↑	#4↑	#5↑
SoTA Specialist	27.53	38.15	36.60	37.87	25.32	25.75	33.86	56.71	58.29	34.27	34.16	36.75	97.50	54.60
VidAgent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LM4LV	24.17	33.05	0.00	0.00	22.28	22.46	28.69	50.12	52.33	0.00	0.00	29.64	0.00	0.00
NExT-GPT-V1.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

C.8 Results of Audio-related Tasks

Audio Comprehension Results. The complete results of all models on audio comprehension are presented in Table 60 and Table 61.

Table 60: Results on Audio Comprehension Group, from #A-C-1 to #A-C-4.

Model	#A-C-1 (Acnt Analy)				#A-C-2 (Ctnt Analy)			#A-C-3 (SpeechEmo Analy)	#A-C-4 (Music Analy)			
	#1↑	#2↑	#3↑	#4↑	#1↑	#2↑	#3↑	#1↑	#1↑	#2↑	#3↑	#4↑
SoTA Specialist	75.20	85.60	90.33	97.95	98.76	95.29	43.20	70.62	83.50	69.3	74.00	89.20
Qwen-Audio-Chat	56.70	75.00	55.40	40.60	93.00	76.80	36.50	**76.80**	61.20	55.20	33.00	1.40
Qwen2-Audio-Instruct	45.80	**99.20**	53.20	92.40	94.40	82.80	**47.20**	61.40	75.20	47.80	19.40	4.80
GAMA	23.40	**86.50**	32.40	85.70	80.90	77.50	34.20	68.00	63.50	63.90	**85.40**	0.00
Pengi	21.40	78.90	35.00	76.20	78.00	65.70	36.50	56.70	61.90	39.20	46.00	0.00
SALMONN-7B	24.70	80.50	85.00	65.10	65.40	58.30	24.10	45.56	60.50	16.20	33.70	0.00
SALMONN-13B	26.04	84.20	**93.50**	67.80	68.70	68.90	31.40	67.80	63.80	19.40	34.60	0.00
WavLLM	37.45	70.15	90.00	60.20	35.60	63.10	24.50	**71.20**	56.80	49.80	12.50	1.20
NExT-GPT-V1.5	12.50	55.40	40.50	64.50	30.50	36.80	20.10	65.80	56.80	43.70	5.50	0.80
PandaGPT (13B)	6.50	58.30	57.30	45.10	24.50	26.40	9.80	45.20	32.40	45.60	5.40	0.50
ImageBind-LLM	10.50	65.00	25.40	50.10	40.70	38.50	18.50	56.80	42.30	25.70	10.40	0.00
ModaVerse-7b-v0	10.50	50.40	35.20	40.30	18.70	20.10	10.30	32.80	24.00	32.60	4.20	0.00
Any-GPT	34.60	62.10	15.00	66.30	37.40	46.10	12.90	63.40	68.40	**78.90**	45.00	0.00
Unified-io-2-XXL	22.50	45.90	13.50	38.70	34.50	32.50	15.80	56.10	56.70	36.80	20.10	0.70

Table 61: Results on Audio Comprehension Group, from #A-C-5 to #A-C-9.

Model	#A-C-5 (Aud-Tech Analy)			#A-C-6 (Aud Analy)		#A-C-7 (Aud QA)		#A-C-8 (Animal-Sound Det)		#A-C-9 (Envir-Sound Det)		
	#1↑	#2↑	#3↑	#1↑	#2↑	#1↑	#2↑	#1↑	#2↑	#1↑	#2↑	#3↑
SoTA Specialist	69.40	80.30	65.90	70.50	55.30	38.90	78.50	77.60	78.20	76.40	86.70	71.10
Qwen-Audio-Chat	63.40	58.40	21.34	19.32	20.25	31.27	**81.60**	**86.30**	**84.00**	**86.11**	**92.60**	56.80
Qwen2-Audio-Instruct	65.70	61.40	10.36	16.82	10.08	34.56	**88.80**	**87.50**	70.40	**71.58**	**86.80**	45.60
GAMA	23.50	25.40	6.40	25.40	28.50	23.60	74.10	**89.60**	**81.50**	72.30	80.50	32.60
Pengi	38.40	15.70	5.20	16.80	22.30	15.40	70.50	**83.60**	71.20	68.40	78.40	36.70
SALMONN-7B	31.70	17.90	5.10	15.40	15.60	17.80	60.40	75.80	65.00	57.80	66.40	24.60
SALMONN-13B	32.30	35.40	6.30	21.00	17.72	19.00	68.90	**79.60**	73.50	60.60	75.90	33.50
WavLLM	53.60	31.40	8.90	29.70	23.40	13.50	78.00	**86.40**	36.40	74.30	77.80	41.60
NExT-GPT-V1.5	24.60	16.50	2.30	16.80	34.50	25.60	70.30	76.90	63.50	69.50	80.90	57.90
PandaGPT-13B	21.50	3.60	0.30	10.50	30.50	15.30	69.20	52.90	56.70	61.50	80.10	55.90
ImageBind-LLM	26.70	14.30	2.00	14.30	22.50	23.40	67.30	63.40	59.20	36.50	72.60	57.10
ModaVerse-7b-v0	14.20	4.10	1.50	2.50	15.30	13.80	56.30	46.80	51.60	58.70	75.70	46.00
Any-GPT	28.40	10.80	9.60	36.10	36.70	28.90	76.40	62.10	73.80	45.20	52.30	36.40
Unified-io-2-XXL	33.10	5.40	7.90	41.30	35.40	12.30	65.10	57.30	69.70	56.80	79.40	45.70

Audio Generation Results. The complete results of all models on audio generation are presented in Table 62 and Table 63.

Table 62: Results on Audio Generation Group, from #A-G-1 to #A-G-7.

Model	#A-G-1 (Audio Edit)	#A-G-2 (Dialog Gen)	#A-G-3 (EmoSpeech Gen)		#A-G-4 (TTS)		#A-G-5 (Txt2Aud)		#A-G-6 (Img2Aud Gen)	#A-G-7 (V2A)
	#1↑	#1↑	#1↓	#2↑	#1↓	#2↑	#1↑	#2↑	#1↑	#1↓
SoTA Specialist	31.50	3.82	4.12	3.15	5.60	3.76	47.04	36.03	51.40	11.52
LLaMA-Omni	0.00	3.01	5.61	2.36	20.15	0.00	0.00	0.00	0.00	∞
Unified-io 2	18.36	2.03	7.86	2.35	78.50	2.54	21.78	11.03	24.31	16.97
Any-GPT	23.50	3.24	6.98	2.15	65.80	1.35	16.52	10.24	14.05	27.49
Next-GPT-V1.5	13.60	1.15	6.78	1.35	100.00	1.02	**53.68**	15.34	1.35	12.36
AudioGPT	0.50	1.32	5.32	**3.89**	45.20	0.00	48.63	10.32	0.00	∞
SpeechGPT	0.10	2.79	5.74	3.14	63.70	0.00	0.00	0.00	0.00	∞
ModaVerse	12.30	1.15	7.52	1.05	100.00	1.00	**50.33**	7.65	1.05	16.45

Table 63: Results on Audio Generation Group, from #A-G-8 to #A-G-11.

Model	#A-G-8 (Style Trans)	#A-G-9 (Speech Trans)			#A-G-10 (Music Gen)				#A-G-11 (Music Trans)	
	#1↓	#1↓	#2↓	#3↓	#1↑	#2↑	#3↑	#4↓	#1↑	#2↑
SoTA Specialist	6.80	7.10	7.70	10.20	25.80	59.01	2.85	3.87	28.16	12.50
LLaMA-Omni	45.30	89.36	93.56	100.00	0.00	0.00	0.00	∞	0.00	0.00
Unified-io 2	86.23	93.21	100.00	90.36	0.00	0.00	0.00	∞	3.15	1.32
Any-GPT	45.36	56.89	95.45	99.34	0.00	0.00	0.00	∞	1.28	3.65
Next-GPT	96.70	99.30	98.40	100.00	0.00	0.00	0.00	∞	8.76	6.78
AudioGPT	46.30	45.68	94.25	100.00	0.00	0.00	0.00	∞	0.00	0.00
SpeechGPT	30.24	57.96	98.67	100.00	0.00	0.00	0.00	∞	0.00	0.00
ModaVerse	100.00	100.00	100.00	100.00	0.00	0.00	0.00	∞	3.59	4.75

C.9 Results of 3D-related Tasks

3D Comprehension Results. The complete results of all models on 3D comprehension are presented in Table 64 to Table 66.

Table 64: Results on 3D Comprehension Group, from #D-C-1 to #D-C-4.

Model	#D-C-1 (3D-Human Cls)					#D-C-2 (3D-Struct Cls)		#D-C-3 (Tech Cls)		#D-C-4 (Indoor-Scene Seg)
	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#2↑	#1↑	#2↑	#1↑
SoTA Specialist	91.18	96.25	94.29	99.50	100.00	99.20	97.50	95.56	100.00	78.50
3D-VisTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PointLLM-7B	6.36	59.28	45.55	48.12	71.50	0.00	15.00	65.71	80.00	0.00
PointLLM-13B	5.90	56.42	52.77	49.87	79.00	5.00	15.00	69.28	87.00	0.00
3D-LLM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AvatarGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 65: Results on 3D Comprehension Group, from #D-C-5 to #D-C-9.

Model	#D-C-5 (Outdoor-Scene Seg)	#D-C-6 (Indoor-Inst Seg)	#D-C-7 (Pose Est)	#D-C-8 (Part Seg)						#D-C-9 (3D Track)
	#1↑	#1↑	#1↓	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#1↑
SoTA Specialist	70.02	81.20	55.00	87.31	93.40	88.62	89.38	81.44	89.52	75.20
3D-VisTA	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PointLLM-7B	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PointLLM-13B	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3D-LLM	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AvatarGPT	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 66: Results on 3D Comprehension Group, from #D-C-10 to #D-C-13.

Model	#D-C-10 (3D-Geo Analy)	#D-C-11 (3D Det)	#D-C-12 (3D QA)							#D-C-13 (3D-Motion Analy)
	#1↓	#1↑	#1↑	#2↑	#1↑	#2↑	#1↑	#2↑	#1↑	#1↑
SoTA Specialist	9.96	68.52	12.40	35.60	67.20	48.50	71.40	49.10	45.80	22.30
3D-VisTA	∞	0.00	**16.00**	34.80	63.30	45.40	69.80	47.20	**48.10**	0.00
PointLLM-7B	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PointLLM-13B	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3D-LLM	∞	0.00	12.00	**36.50**	65.60	47.20	68.80	48.00	46.30	0.00
AvatarGPT	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.70

3D Generation Results. The complete results of all models on 3D generation are presented in Table 67 and Table 68.

Table 67: Results on 3D Generation Group, from #D-G-1 to #D-G-4.

Model	#D-G-1 (PC Complt)	#D-G-2 (PC2M Recon)		#D-G-3 (Txt2PC Gen)				#D-G-4 (Txt2M Gen)			
	#1↓	#1↓	#2↓	#1↑	#2↑	#3↑	#4↑	#1↑	#2↑	#3↑	#4↑
SoTA Specialist	0.22	9.32E-05	4.93E-05	24.94	25.10	24.07	23.56	26.42	25.22	25.93	25.18
MotionGPT-1	∞	∞	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-2	∞	∞	∞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaMA-Mesh	∞	∞	∞	0.00	0.00	0.00	0.00	20.06	14.43	18.06	17.65

Table 68: Results on 3D Generation Group, from #D-G-5 to #D-G-9.

Model	#D-G-5 (Img2PC Gen)				#D-G-6 (Img2M Gen)				#D-G-7 (RGBD2PC Recon)	#D-G-8 (RGBD2Mesh Recon)	#D-G-9 (Txt2Motion Gen)
	#1↑	#2↑	#3↑	#4↑	#1↑	#2↑	#3↑	#4↑	#1↓	#1↓	#1↓
SoTA Specialist	77.06	78.27	78.87	78.04	83.30	82.88	84.43	83.96	6540.02	6540.02	0.23
MotionGPT-1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	0.51
MotionGPT-2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	0.60
LLaMA-Mesh	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	∞	∞	∞

C.10 Results of NLP Tasks

All the results of all generalists on NLP tasks are shown in Table 69, Table 71, Table 73, Table 75, Table 77, Table 79, Table 81, Table 83, Table 85, Table 87, and Table 89.

Table 69: Results on NLP Group, #L-1.

Model	#L-1 (Cog QA)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
SoTA Specialist	87.33	95.00	31.47	28.42	80.00	33.80	40.75	86.45	80.39
Meta-Llama-3.1-8B-Instruct	74.90	82.40	13.11	5.18	65.80	6.89	16.26	43.00	50.20
Qwen2.5-7B-Instruct	74.71	82.00	16.36	4.46	67.80	6.30	19.38	53.65	37.99
Gemma-2-9b-it	80.59	78.60	20.79	8.08	75.60	9.28	17.19	29.19	59.65
ChatGLM-6b	52.75	67.40	8.27	2.69	43.80	13.03	5.57	43.00	24.21
Vicuna-7b-v1.5	44.51	63.80	0.35	1.01	46.40	0.47	0.08	44.18	22.24
InternLM-Chat-7b	78.24	60.20	13.40	2.53	49.20	9.96	6.79	43.20	25.20
GPT-J-6B	20.39	47.00	0.50	0.23	38.60	0.32	0.52	51.28	27.95
Falcon3-7B-Instruct	74.31	78.60	18.71	4.24	51.20	7.27	11.83	57.79	27.17
Baichuan2-7B-Base	20.98	47.80	0.07	1.11	40.80	0.12	1.29	54.64	25.20
Ministral-8B-Instruct-2410	71.76	80.20	22.14	19.35	65.00	21.33	20.19	37.48	38.19
Yi-Ligntning	79.22	84.60	13.31	6.30	74.20	7.42	22.78	54.44	33.27
GPT-3.5-turbo	20.78	47.40	12.83	3.32	41.00	3.11	23.15	55.42	26.18
GPT-4v	20.78	47.40	14.62	2.26	41.00	2.87	37.43	55.42	26.18
GPT-4o	20.78	47.40	16.00	2.92	41.00	3.20	23.34	55.42	26.18
GPT4o-mini	20.78	47.40	14.04	2.04	41.00	2.72	18.71	55.42	26.18
GPT-4o-4096	20.78	47.40	17.68	3.77	41.00	3.89	23.28	55.42	26.18
ChatGPT-4o-latest	20.78	47.40	15.33	3.57	41.00	3.35	21.56	55.42	26.18
Claude-3.5-Sonnet	64.71	73.20	22.28	7.37	56.20	22.85	20.34	54.24	43.90
Claude-3.5-Opus	60.78	75.60	17.11	5.91	50.00	15.25	25.56	44.18	38.58
Emu2-32B	56.27	64.80	14.94	3.58	47.20	11.24	17.72	44.18	36.22
DetGPT	51.18	62.80	12.78	0.09	42.40	4.18	10.66	38.46	31.10
InternVL2.5-8B	78.43	77.60	21.57	8.00	65.40	15.71	20.55	44.18	54.92
InternVL2.5-4B	81.37	78.60	20.75	6.41	61.60	20.25	19.66	34.71	50.00
NExT-GPT-V1.5	43.26	56.84	0.35	0.76	40.58	0.39	0.08	43.65	0.00
InternVL2.5-2B	72.35	76.00	16.36	2.49	43.00	11.23	14.74	25.83	33.07
Monkey-10B-chat	26.27	52.20	2.46	15.99	37.80	3.05	5.97	41.42	24.01
DeepSeek-VL-7B	72.54	80.60	11.05	3.70	14.20	7.64	14.20	35.89	29.92
Qwen2-VL-7B	36.00	60.00	11.63	3.97	33.00	6.63	12.00	19.00	33.00
Qwen-VL-Chat	59.80	63.80	0.29	0.92	47.40	0.42	0.64	36.88	26.97
Qwen-Audio-Chat	59.80	56.40	9.89	3.04	48.60	6.08	6.05	41.42	23.62
Qwen2-Audio-Instruct	85.00	82.00	4.77	2.87	0.00	5.58	19.43	60.00	0.00
MoE-LLAVA-Phi2-2.7B-4e-384	64.71	71.80	13.75	4.24	53.80	8.76	9.78	35.50	26.77
mPLUG-Owl2-LLaMA2-7b	61.96	73.40	11.24	2.26	58.00	13.03	6.52	41.42	25.59
Phi-3.5-Vision-Instruct	74.71	82.00	19.05	5.06	62.40	19.57	21.83	40.83	29.92
Cambrian-1-8B	21.76	49.20	4.95	2.57	41.00	6.13	8.56	55.42	26.38
MiniGPT4-LLaMA2	72.55	47.40	0.35	1.22	63.60	0.44	0.00	37.28	30.71
InternVL-Chat-V1-5	73.13	68.20	14.61	5.80	59.40	7.43	7.46	39.84	35.43
Mini-InternVL-Chat-4B-V1-5	60.39	75.60	15.56	5.50	53.80	12.21	14.69	40.03	27.75
InternLM-XComposer2-VL-1.8B	65.88	76.40	15.13	2.33	46.60	6.33	6.98	31.75	22.04
GPT4RoI	21.56	47.60	1.27	2.77	41.20	2.50	0.72	55.42	26.18
GLaMM	20.78	47.40	0.00	0.00	41.00	1.28	4.77	55.42	26.18
LLaVA-NeXT-13B	37.25	62.60	9.47	3.64	43.60	6.05	6.63	42.80	27.56
LLaVA-NeXT-34B	43.92	64.60	17.93	5.88	47.20	7.71	8.66	40.83	28.74
Pixtral-12B	46.67	72.40	18.95	6.25	47.80	6.43	7.10	42.21	29.13
SEED-LLaMA-13B	15.29	46.60	6.67	1.03	19.20	2.76	5.43	41.42	24.61
BLIP2	43.13	48.60	11.41	1.08	3.40	4.62	13.49	39.05	32.09
MiniMonkey	70.78	76.20	17.02	2.31	51.00	7.10	17.49	37.48	27.76
DeepSeek-VL-7B	46.86	0.00	5.00	6.84	44.20	6.93	0.00	46.35	26.18
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Model	#L-1 (Cog QA)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
CogVLM-Chat	55.88	69.20	14.16	6.53	53.00	13.59	19.73	44.38	35.43
ShareGPT4V-7B	47.45	62.00	10.76	4.26	42.00	8.30	17.69	41.62	27.76
ShareGPT4V-13B	54.51	67.00	15.50	5.37	50.40	9.17	18.78	39.64	35.24
BLIP-3 (XGen-MM)	51.57	62.80	16.86	5.79	49.20	10.97	17.94	40.24	36.61
AnyGPT	21.96	24.00	6.33	4.84	17.00	6.16	7.87	19.92	17.91
MiniCPM3-4B	52.35	72.80	18.91	6.17	55.80	16.15	20.25	47.93	37.01
LaVIT-V2 (7B)	39.41	60.20	12.92	11.42	45.20	14.59	14.48	37.28	28.94
GLM-VL-Chat	55.69	71.40	16.78	5.44	55.60	14.07	20.71	49.51	39.96
Gemini-1.5-Pro	81.76	83.40	20.86	7.25	62.80	21.03	22.85	54.24	53.15
Gemini-1.5-Flash	74.51	76.60	19.27	6.63	56.00	18.49	17.74	44.18	46.26
OMG-LLaVA-InternLM20B	49.02	47.20	13.64	2.83	41.20	9.21	13.07	36.65	27.59
Idefics3-8B-Llama3	78.04	76.40	16.71	6.39	57.80	16.36	18.20	43.00	41.54
Yi-Vision-v2	20.78	47.40	16.18	6.34	41.00	2.82	22.19	55.42	26.18
Qwen2-VL-72B	41.33	63.95	20.84	8.31	45.77	9.65	25.04	56.93	27.34
Otter	30.00	49.10	1.78	0.76	40.80	0.76	1.01	54.83	26.57
Show-o	18.00	45.00	1.17	8.62	20.00	7.19	12.64	29.10	31.00
NExT-Chat	20.78	47.40	2.88	1.19	41.00	1.04	3.52	55.42	26.18
InternVL2-26B	81.70	68.00	18.40	6.50	63.80	10.70	19.20	42.90	33.80
Qwen2-VL-72B	81.50	86.60	24.80	14.10	76.00	29.50	22.20	50.60	73.40
DeepSeek-VL-2-small	74.30	80.20	8.90	4.20	57.20	6.20	12.00	36.70	28.00
DeepSeek-VL-2	77.80	79.00	11.10	3.70	65.60	6.90	16.70	46.20	28.70
LLaVA-One-Vision-7B	76.50	79.00	19.90	5.60	64.60	15.30	18.80	42.80	46.30
LLaVA-One-Vision-72B	85.90	87.40	27.80	6.90	76.80	25.90	26.00	53.50	64.00
Sa2VA-8B	84.90	82.20	25.60	14.60	71.40	18.60	22.50	49.10	59.10
Sa2VA-26B	76.50	79.80	20.60	7.30	75.40	22.10	13.40	45.40	68.90
CoLVA-2B	49.60	68.80	11.00	2.90	42.20	9.40	12.10	27.60	20.90
CoLVA-4B	84.90	84.80	16.70	5.00	69.40	11.20	16.40	39.80	50.40
Long-LLaVA	26.90	49.60	13.30	2.50	41.20	10.90	13.30	55.40	25.40
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	40.32	56.84	3.50	1.67	35.42	2.56	1.09	40.37	0.00
PandaGPT (13B)	15.78	67.54	0.14	0.90	0.00	0.45	0.00	47.32	0.00
AnyGPT	21.96	24.00	6.33	4.84	17.00	6.16	7.87	19.92	17.91
GAMA	34.60	57.80	4.75	2.29	0.00	4.85	0.35	76.50	0.00
Pengi	20.40	40.80	2.56	1.89	3.45	3.24	0.23	56.80	0.00
SALMONN-7B	23.53	70.90	1.84	1.03	0.00	1.48	0.61	68.23	0.00
SALMONN-13B	24.31	71.20	1.78	0.90	23.40	2.13	0.65	69.60	0.00
WavLLM	32.50	57.46	4.32	2.01	0.00	4.32	0.21	83.10	0.00
ImageBind-LLM	28.70	56.71	3.45	1.78	0.00	2.56	0.10	58.30	0.00
Unified-io-2-XXL	21.56	47.60	1.27	2.77	41.20	2.50	0.72	55.42	26.18
ModaVerse-7b-v0	15.78	37.23	3.10	0.50	0.00	0.34	0.00	43.24	0.00
AudioGPT-GPT4	21.45	42.34	10.23	2.04	39.80	2.68	38.56	54.98	25.38
SpeechGPT-7B-com	38.60	54.23	4.23	1.99	0.00	4.09	0.12	45.70	0.00
LLaMA-Omni	56.67	63.50	13.11	5.18	15.60	6.89	16.26	43.00	0.00
3D-LLM	20.78	47.40	17.55	8.61	41.00	21.31	7.92	55.42	26.18
PointLLM-7B	20.78	47.40	0.35	1.15	41.00	0.48	0.87	55.42	26.18
PointLLM-13B	20.20	47.40	0.35	1.20	40.80	0.47	0.88	55.42	26.77
3D-VisTA	20.78	47.40	2.48	0.28	41.00	0.56	3.65	55.42	26.18
MotionGPT-T5	20.78	47.40	0.27	0.11	41.00	0.04	5.52	55.42	26.18
MotionGPT-LLaMA	23.14	47.40	0.98	1.14	36.00	0.81	0.90	50.10	23.43
AvatarGPT	20.78	47.40	2.82	0.30	41.00	0.58	4.17	55.42	26.18
LLaMA-Mesh	65.69	72.40	0.35	1.64	57.20	0.48	0.85	39.45	25.98

Table 71: Results on NLP Group, #L-2.

Model	#L-2 (Ethics NLP)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
SoTA Specialist	95.62	45.39	75.00	79.80	94.20	95.40	99.88	95.80	95.00
Meta-Llama-3.1-8B-Instruct	83.40	24.08	48.80	67.00	56.00	81.00	45.40	94.20	11.00
Qwen2.5-7B-Instruct	84.80	32.24	36.60	46.80	80.40	66.40	91.40	55.60	62.80
Gemma-2-9b-it	89.60	35.31	35.00	57.40	86.60	25.40	95.00	58.60	54.80
ChatGLM-6b	68.40	28.98	31.80	0.00	1.60	6.60	67.20	51.80	42.80
Vicuna-7b-v1.5	50.40	20.44	29.80	0.00	0.00	0.00	0.00	0.00	0.00
InternLM-Chat-7b	75.60	26.83	36.00	43.80	9.40	25.80	27.80	29.00	6.00
GPT-J-6B	72.00	3.76	21.00	0.00	0.00	0.00	0.00	0.00	0.00
Falcon3-7B-Instruct	87.20	31.23	44.20	49.20	82.60	68.00	40.40	54.80	67.60
Baichuan2-7B-Base	82.40	4.46	20.20	0.00	0.00	0.00	0.00	0.00	0.00
Ministral-8B-Instruct-2410	87.40	21.66	42.80	48.20	79.60	40.40	88.40	41.80	37.60
Yi-Ligntning	91.40	34.43	53.40	55.60	74.60	42.60	96.40	45.20	51.20
GPT-3.5-turbo	84.40	26.35	20.20	56.40	78.40	47.00	69.00	48.80	76.20
GPT-4v	84.40	29.63	20.20	54.00	82.60	77.40	99.60	64.00	49.80
GPT-4o	84.40	30.29	20.20	54.20	80.40	75.60	99.00	63.60	55.40
GPT4o-mini	84.40	29.33	20.20	55.00	80.20	63.60	98.80	59.60	59.00
GPT-4o-4096	84.40	30.83	20.20	55.20	84.20	55.80	99.20	65.40	66.20
ChatGPT-4o-latest	84.40	27.61	20.20	53.40	82.00	71.00	99.20	61.80	68.00
Claude-3.5-Sonnet	85.40	39.21	38.80	39.75	55.21	45.19	59.26	57.57	60.07
Claude-3.5-Opus	86.40	34.19	35.40	43.75	43.99	38.95	55.59	57.13	49.76
Emu2-32B	79.60	29.15	36.00	34.44	46.40	37.46	48.13	53.04	44.61
DetGPT	73.40	18.91	28.80	29.53	36.38	31.68	45.03	45.16	44.91
InternVL2.5-8B	85.20	24.83	45.60	22.99	23.17	33.21	35.42	80.80	78.58
InternVL2.5-4B	78.80	18.69	39.00	20.49	20.76	33.75	32.58	80.80	78.98
NExT-GPT-V1.5	50.78	20.44	30.74	23.10	18.60	10.98	21.43	25.70	0.00
InternVL2.5-2B	70.60	23.56	21.88	21.88	17.97	31.47	34.11	64.80	65.25
Monkey-10B-chat	69.00	5.13	23.20	5.35	5.97	12.17	8.54	43.20	43.63
DeepSeek-VL-7B	85.00	29.48	35.80	18.12	26.04	33.82	33.59	68.00	69.69
Qwen2-VL-7B	27.00	25.58	29.00	46.00	34.00	1.00	26.00	47.00	12.00
Qwen-VL-Chat	82.00	9.65	32.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-Audio-Chat	49.00	32.49	31.40	46.80	43.60	68.30	60.40	31.50	7.80
Qwen2-Audio-Instruct	66.85	33.24	0.00	47.00	49.60	68.20	61.60	30.00	6.20
MoE-LLAVA-Phi2-2.7B-4e-384	79.20	30.48	28.40	52.20	80.20	13.00	39.60	47.00	63.40
mPLUG-Owl2-LLaMA2-7b	82.40	31.59	35.20	53.60	78.40	46.40	76.40	48.60	29.80
Phi-3.5-Vision-Instruct	83.40	30.88	41.40	50.60	83.20	91.20	89.20	60.60	55.20
Cambrian-1-8B	84.40	22.09	20.40	44.20	1.80	3.80	0.20	0.00	0.80
MiniGPT4-LLaMA2	81.60	16.20	43.80	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-Chat-V1-5	87.80	27.53	39.20	22.83	26.50	34.19	32.93	75.00	69.09
Mini-InternVL-Chat-4B-V1-5	71.80	22.41	35.80	19.28	23.91	32.03	31.11	74.20	74.34
InternLM-XComposer2-VL-1.8B	75.20	29.08	32.20	18.52	8.52	32.73	26.87	59.60	68.88
GPT4RoI	84.20	19.25	20.40	24.20	20.29	12.94	21.43	47.20	33.13
GLaMM	84.40	15.21	20.20	8.33	9.58	12.72	12.72	48.20	33.13
LLaVA-NeXT-13B	75.60	27.81	24.80	53.20	56.80	36.20	64.40	49.60	53.20
LLaVA-NeXT-34B	78.60	36.41	28.20	49.20	62.80	44.80	71.20	51.20	58.40
Pixtral-12B	80.20	38.20	34.40	52.40	61.40	40.20	74.20	52.00	59.60
SEED-LLaMA-13B	44.20	14.32	14.60	38.80	47.60	17.40	45.60	39.80	30.60
BLIP2	76.80	43.59	40.60	49.00	60.80	38.60	62.80	43.60	57.20
MiniMonkey	72.40	35.01	34.00	54.80	56.60	13.60	71.20	48.40	78.00
DeepSeek-VL-7B	76.40	31.40	22.80	22.28	22.42	28.05	0.00	52.40	47.68
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	83.20	24.30	33.60	34.64	44.71	41.14	47.47	48.75	44.84
ShareGPT4V-7B	72.60	21.24	30.80	28.76	36.28	30.66	44.28	47.46	40.08
ShareGPT4V-13B	78.00	25.67	34.00	36.74	45.58	33.09	43.25	51.01	42.50

Model	#L-2 (Ethics NLP)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
BLIP-3 (XGen-MM)	78.80	25.63	36.00	35.55	41.23	38.53	45.28	48.25	47.43
AnyGPT	29.20	9.42	12.80	14.54	16.12	15.31	17.22	19.70	15.66
MiniCPM3-4B	86.60	26.23	32.20	33.08	46.01	33.86	48.91	58.84	52.47
LaVIT-V2 (7B)	64.40	21.72	27.80	26.72	33.86	26.62	36.15	44.13	43.77
GLM-VL-Chat	86.80	27.19	30.40	32.97	43.97	32.48	49.16	57.70	51.78
Gemini-1.5-Pro	87.00	34.33	44.20	51.60	69.80	69.00	81.20	62.60	78.40
Gemini-1.5-Flash	83.40	29.98	40.60	47.00	60.80	63.60	77.40	58.00	75.20
OMG-LLaVA-InternLM20B	59.80	12.26	21.20	19.20	17.00	9.40	24.20	20.20	32.60
Idefics3-8B-Llama3	82.20	27.71	39.60	48.20	63.60	61.20	76.40	59.80	73.20
Yi-Vision-v2	84.40	20.63	20.20	53.60	67.60	41.99	96.20	58.80	48.60
Qwen2-VL-72B	89.20	25.73	20.27	46.67	74.67	80.02	96.00	47.33	62.67
Otter	69.00	10.33	20.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	79.00	6.89	23.00	26.73	17.75	39.24	31.10	36.87	21.54
NExT-Chat	20.78	7.15	20.20	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2-26B	90.00	29.10	43.40	51.20	81.50	41.40	91.40	53.70	64.00
Qwen2-VL-72B	91.80	32.80	59.40	52.40	73.40	77.40	93.40	54.20	62.20
DeepSeek-VL-2-small	83.80	34.20	39.20	53.60	65.60	24.80	68.40	26.40	36.20
DeepSeek-VL-2	86.80	35.10	42.20	56.60	81.20	65.80	78.80	28.40	69.00
LLaVA-One-Vision-7B	89.20	25.30	49.00	54.80	82.80	58.80	90.20	56.80	61.60
LLaVA-One-Vision-72B	92.40	24.70	58.00	0.20	0.00	44.80	99.40	54.60	3.80
Sa2VA-8B	89.20	30.40	48.80	48.60	84.40	38.20	76.20	39.80	62.80
Sa2VA-26B	91.60	28.90	52.60	51.00	85.40	41.60	98.80	54.80	71.20
CoLVA-2B	67.40	29.00	30.20	51.20	77.40	1.80	6.20	55.80	34.40
CoLVA-4B	81.40	30.50	44.40	54.00	71.60	28.60	93.80	36.40	58.80
Long-LLaVA	87.20	30.80	22.00	36.60	45.20	64.40	66.80	60.20	32.20
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	57.83	10.32	23.45	13.50	38.40	10.98	23.45	27.50	0.05
PandaGPT (13B)	60.40	19.32	0.00	1.20	2.10	0.00	0.00	0.00	0.00
AnyGPT	29.20	9.42	12.80	14.54	16.12	15.31	17.22	19.70	15.66
GAMA	35.87	17.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pengi	68.51	14.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SALMONN-7B	1.00	21.59	0.00	1.60	4.40	1.60	1.00	1.00	1.60
SALMONN-13B	1.20	22.67	0.00	1.50	4.70	1.40	1.10	0.00	0.00
WavLLM	25.83	14.56	0.00	1.40	4.30	2.40	1.30	0.00	0.00
ImageBind-LLM	56.40	13.60	0.00	1.10	4.50	1.70	1.10	0.00	0.00
Unified-io-2-XXL	56.80	17.56	17.80	16.50	18.50	10.30	11.40	4.30	23.60
ModaVerse-7b-v0	49.39	15.80	0.00	1.02	1.90	0.00	0.00	0.00	0.00
AudioGPT-GPT4	84.30	30.40	20.10	54.10	81.90	77.40	98.70	63.60	49.70
SpeechGPT-7B-com	45.60	13.56	0.00	1.20	4.10	0.00	0.00	0.00	0.00
LLaMA-Omni	44.50	24.08	0.00	45.30	34.20	45.30	23.10	14.50	11.00
3D-LLM	84.40	9.29	20.20	1.20	18.00	0.00	0.00	0.00	0.20
PointLLM-7B	84.40	19.91	20.20	0.00	0.00	0.00	0.00	0.00	0.00
PointLLM-13B	84.60	19.69	23.60	0.00	0.00	0.00	0.00	0.00	0.00
3D-VisTA	84.40	0.64	20.20	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-T5	84.40	3.56	20.20	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-LLaMA	71.20	5.30	28.40	0.00	0.00	0.00	0.00	0.00	0.00
AvatarGPT	84.40	11.48	20.20	0.00	0.00	0.00	0.00	0.00	0.00
LLaMA-Mesh	87.20	17.93	45.20	0.00	0.00	0.00	0.00	0.00	0.00

Table 73: Results on NLP Group, from #L-3 to #L-4.

Model	#L-3 (Domain QA)							#L-4 (Social QA)					
	#1↑	#2↑	#3↑	#4↑	#5↑	#↑6	#7↑	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑
SoTA Specialist	82.62	48.79	73.55	53.31	94.50	96.67	88.00	48.73	82.00	49.21	97.30	91.00	57.74
Meta-Llama-3.1-8B-Instruct	29.20	22.89	36.90	41.42	83.00	84.24	81.80	42.66	63.80	29.39	94.40	93.00	39.89
Qwen2.5-7B-Instruct	20.56	27.48	41.47	27.69	86.80	83.43	77.20	35.58	63.00	20.48	94.20	93.80	22.69
Gemma-2-9b-it	15.78	28.87	37.97	15.70	89.80	80.40	77.20	23.13	62.60	14.85	93.40	93.80	13.42
ChatGLM-6b	19.60	23.32	30.54	27.28	52.60	55.76	51.60	31.83	57.60	22.98	68.60	71.00	24.59
Vicuna-7b-v1.5	15.63	23.22	15.62	14.90	59.00	48.08	57.60	18.83	55.40	19.06	67.40	71.00	15.46
InternLM-Chat-7b	14.20	15.81	29.55	20.55	67.20	73.13	70.80	23.91	67.80	17.15	84.60	81.60	15.51
GPT-J-6B	12.50	12.93	6.11	10.87	43.40	32.12	36.60	7.86	48.20	10.40	36.20	31.60	12.42
Falcon3-7B-Instruct	24.60	26.11	38.93	38.45	81.20	65.45	74.60	41.65	65.20	27.99	87.80	87.80	30.33
Baichuan2-7B-Base	12.63	14.51	8.38	12.63	48.20	33.13	39.20	9.04	51.40	9.78	35.40	30.00	13.65
Ministral-8B-Instruct-2410	16.39	27.41	38.60	21.27	82.60	83.84	76.60	28.74	66.60	14.68	92.20	92.40	16.87
Yi-Ligntning	27.08	26.62	41.18	37.58	89.60	85.45	80.20	42.34	68.20	28.95	95.00	94.00	34.55
GPT-3.5-turbo	14.22	29.65	36.26	24.38	48.20	33.13	40.00	28.14	51.20	14.95	35.60	30.00	13.58
GPT-4v	22.56	25.12	36.75	36.22	48.20	33.13	40.00	37.08	51.20	25.99	25.60	30.00	25.40
GPT-4o	19.95	26.31	43.04	27.21	48.20	33.13	40.00	33.32	51.20	19.45	35.60	30.00	19.43
GPT4o-mini	21.42	28.98	41.72	33.83	48.20	33.13	40.00	38.67	51.20	23.10	35.60	30.00	24.83
GPT-4o-4096	19.36	27.90	42.29	27.30	48.20	33.13	40.00	33.38	51.20	20.36	35.60	30.00	19.12
ChatGPT-4o-latest	19.67	28.22	43.00	26.80	48.20	33.13	40.00	32.25	51.20	18.64	35.60	30.00	17.50
Claude-3.5-Sonnet	47.83	37.75	53.34	37.66	69.60	76.20	54.40	41.68	60.20	30.02	58.60	61.00	42.93
Claude-3.5-Opus	46.29	33.32	43.88	28.87	69.40	71.20	49.60	34.60	57.00	24.78	59.60	51.00	33.36
Emu2-32B	40.15	29.51	46.14	26.70	67.60	69.40	49.80	30.88	48.40	23.51	51.80	49.20	33.57
DetGPT	36.27	21.60	40.13	21.79	61.00	62.80	42.00	28.70	46.00	18.50	43.40	43.00	27.02
InternVL2.5-8B	74.80	37.58	65.80	24.73	92.20	90.60	31.10	24.89	25.92	33.72	26.24	31.72	44.54
InternVL2.5-4B	78.40	34.30	65.00	22.55	91.80	90.60	26.21	26.87	33.41	19.32	38.71	36.48	36.36
NExT-GPT-V1.5	15.46	34.87	10.56	12.50	54.74	46.50	53.20	18.56	53.40	18.40	63.50	69.50	13.68
InternVL2.5-2B	65.80	35.48	61.20	25.94	77.20	75.40	28.97	22.94	33.01	12.06	28.83	31.40	25.34
Monkey-10B-chat	41.80	8.41	47.60	6.09	37.40	43.00	7.65	8.59	11.82	14.87	6.21	11.28	18.78
DeepSeek-VL-7B	69.60	37.17	58.80	25.01	84.00	85.20	29.07	24.30	21.08	18.84	19.38	19.90	18.63
Qwen2-VL-7B	27.22	19.56	24.37	35.59	37.00	67.00	53.00	32.20	42.00	29.45	67.00	69.00	38.75
Qwen-VL-Chat	18.43	14.25	7.97	15.07	56.20	72.20	60.00	11.93	60.60	16.52	72.20	61.60	17.60
Qwen-Audio-Chat	23.20	27.51	35.14	32.37	54.20	58.98	59.00	37.24	51.00	25.99	81.20	74.00	28.59
Qwen2-Audio-Instruct	22.04	26.94	35.01	31.89	65.33	86.60	78.58	37.84	78.65	24.22	90.11	0.00	26.65
MoE-LLAVA-Phi2-2.7B-4e-384	24.06	28.98	34.26	32.01	60.00	71.00	55.40	36.03	59.60	26.12	71.00	66.60	28.89
mPLUG-Owl2-LLaMA2-7b	25.13	23.25	34.04	35.23	71.80	83.60	75.40	39.15	66.40	28.12	83.60	83.00	32.44
Phi-3.5-Vision-Instruct	24.27	26.39	38.03	37.09	79.40	91.00	78.20	40.94	66.60	26.73	91.00	88.00	34.85
Cambrian-1-8B	26.77	17.90	26.20	37.79	48.40	37.00	41.20	28.41	51.20	24.05	37.00	32.20	36.13
MiniGPT4-LLaMA2	26.85	14.42	19.58	0.00	68.20	84.60	73.20	23.81	63.40	0.00	84.60	66.40	32.67
InternVL-Chat-V1-5	68.80	36.29	61.20	25.20	84.60	86.00	28.85	24.68	19.83	16.15	18.50	18.16	25.45
Mini-InternVL-Chat-4B-V1-5	72.00	33.38	61.60	21.77	87.80	86.80	25.55	27.77	26.71	12.11	22.99	25.42	30.81
InternLM-XComposer2-VL-1.8B	57.80	34.44	57.20	22.75	75.80	76.80	25.04	26.40	18.82	8.95	17.57	21.57	16.96
GPT4RoI	40.60	17.49	51.40	23.36	36.40	29.60	26.25	3.25	5.47	6.68	0.90	0.60	0.00
GLaMM	40.00	12.46	51.20	8.70	35.60	30.00	7.40	20.24	9.04	3.07	9.16	9.55	0.00
LLaVA-NeXT-13B	27.85	23.33	36.21	20.22	46.40	55.60	46.60	32.13	60.40	19.74	62.60	51.60	27.55
LLaVA-NeXT-34B	23.78	26.84	33.25	31.89	56.60	58.80	50.20	35.24	57.20	23.19	67.40	61.20	26.64
Pixtral-12B	24.67	24.33	36.94	32.92	62.20	62.00	55.60	40.17	62.40	21.02	69.80	66.80	29.71
SEED-LLaMA-13B	12.05	13.95	24.67	15.48	29.20	47.60	42.80	19.07	38.40	15.16	33.60	24.40	20.49
BLIP2	29.27	23.32	41.83	37.69	64.20	63.00	59.20	46.29	68.40	41.84	79.60	88.40	52.61
MiniMonkey	17.43	15.90	31.21	25.90	62.80	65.86	64.40	32.76	58.00	23.26	80.40	78.40	23.68
DeepSeek-VL-7B	50.80	33.74	56.20	0.00	61.80	58.00	0.00	24.40	15.43	12.02	14.56	17.63	1.53
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	45.61	25.47	46.78	34.10	63.40	67.80	53.20	29.83	49.60	27.09	49.60	48.20	35.70
ShareGPT4V-7B	33.20	24.84	41.18	22.86	62.60	64.20	42.20	27.93	46.40	20.43	46.00	45.60	27.66
ShareGPT4V-13B	36.19	23.32	40.73	26.17	65.40	64.40	47.80	30.45	44.60	22.43	50.40	49.40	27.20

Model	#L-3 (Domain QA)							#L-4 (Social QA)					
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑
BLIP-3 (XGen-MM)	41.17	26.99	44.56	25.79	65.40	68.40	47.60	33.70	46.00	24.40	48.80	45.60	30.89
AnyGPT	13.45	10.84	16.73	10.64	24.80	24.20	18.00	0.00	16.00	7.96	18.60	18.20	10.80
MiniCPM3-4B	51.94	34.47	52.23	31.14	69.80	73.00	45.20	33.55	46.00	25.14	56.40	54.00	37.87
LaVIT-V2 (7B)	40.36	23.12	39.05	25.22	61.60	58.60	36.40	25.57	37.80	23.52	45.20	39.00	26.88
GLM-VL-Chat	50.08	36.02	52.19	31.23	68.00	71.00	43.60	29.72	43.20	25.78	56.40	53.80	37.28
Gemini-1.5-Pro	29.14	34.32	41.77	36.41	89.60	90.20	84.20	42.15	81.40	27.23	86.60	88.40	40.45
Gemini-1.5-Flash	26.93	30.08	37.04	33.83	86.20	84.40	78.20	38.47	72.60	23.78	80.20	82.60	36.83
OMG-LLaVA-InternLM20B	16.07	14.54	26.76	12.95	53.80	51.20	29.80	15.42	14.80	12.91	14.80	16.20	12.86
Idefics3-8B-Llama3	28.54	29.32	35.93	31.89	87.80	88.20	73.60	39.30	74.40	22.05	78.00	79.60	36.13
Yi-Vision-v2	15.03	23.04	37.95	30.74	48.20	33.13	40.00	35.67	51.20	16.25	35.60	30.00	21.46
Qwen2-VL-72B	16.47	25.99	34.25	31.83	69.32	46.24	66.31	34.96	66.34	21.97	47.88	61.92	22.47
Otter	19.58	18.72	9.86	21.66	44.60	38.38	46.00	15.12	47.40	18.77	40.60	34.20	20.50
Show-o	14.13	9.84	13.04	23.23	57.00	38.00	46.00	19.57	38.00	18.65	32.00	28.00	27.79
NExT-Chat	11.62	12.13	6.66	14.17	48.20	33.13	40.00	12.88	51.20	16.85	35.60	30.00	19.15
InternVL2-26B	19.30	19.70	33.40	27.10	81.60	75.90	71.40	32.70	65.00	21.70	86.20	86.40	24.70
Qwen2-VL-72B	22.80	27.80	38.50	36.50	91.20	83.80	79.20	39.50	70.10	25.00	95.40	96.20	28.20
DeepSeek-VL-2-small	21.40	29.20	40.50	29.70	77.00	80.20	81.40	37.40	65.80	26.50	92.80	92.60	28.20
DeepSeek-VL-2	21.90	28.10	42.40	31.80	80.60	82.80	80.80	40.30	67.60	24.70	92.60	95.40	26.80
LLaVA-One-Vision-7B	20.40	21.90	35.30	32.80	82.60	78.20	73.80	33.80	64.80	23.60	88.20	93.00	28.70
LLaVA-One-Vision-72B	23.50	25.00	36.60	39.30	91.20	85.30	81.00	41.10	70.00	28.10	96.40	96.00	35.20
Sa2VA-8B	23.30	26.50	38.20	37.60	84.60	82.60	78.80	40.50	66.80	26.80	92.80	91.40	32.90
Sa2VA-26B	26.20	27.10	38.40	40.00	87.20	77.40	76.20	41.70	68.40	28.40	89.00	89.00	34.80
CoLVA-2B	19.90	20.20	33.50	30.70	53.20	54.30	51.40	34.20	49.80	24.20	66.40	60.60	27.30
CoLVA-4B	21.20	27.80	38.30	30.90	83.60	79.80	80.80	36.20	68.60	23.60	93.00	89.40	26.80
Long-LLaVA	23.10	25.70	30.90	34.80	50.20	36.40	42.60	34.40	54.00	28.10	38.40	46.60	36.20
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	13.47	14.58	15.60	13.20	54.74	45.60	52.30	20.56	49.82	13.40	60.34	62.34	16.83
PandaGPT (13B)	9.78	5.34	17.61	14.90	45.10	22.08	0.00	15.89	10.38	11.15	56.90	24.00	15.46
AnyGPT	13.45	10.84	16.73	10.64	24.80	24.20	18.00	0.00	16.00	7.96	18.60	18.20	10.80
GAMA	22.58	35.84	24.29	28.91	0.00	63.83	65.50	27.30	0.00	21.57	0.00	47.21	25.57
Pengi	14.35	30.46	20.67	26.63	0.00	67.00	45.60	20.40	0.00	20.89	0.00	45.30	15.67
SALMONN-7B	11.63	7.06	28.09	18.58	66.36	32.00	0.00	21.90	20.00	14.68	68.44	28.00	19.19
SALMONN-13B	11.87	7.31	29.78	17.91	76.53	32.13	0.00	22.50	22.10	13.02	77.80	29.70	17.40
WavLLM	21.45	36.79	25.43	29.01	0.00	66.55	57.35	28.09	0.00	22.12	0.00	68.85	26.23
ImageBind-LLM	23.50	45.60	24.30	28.70	0.00	78.00	56.00	29.80	0.00	12.40	0.00	48.75	15.40
Unified-io-2-XXL	16.51	14.35	15.67	12.40	21.34	24.67	20.56	2.50	5.03	6.78	0.80	0.60	0.00
ModaVerse-7b-v0	9.45	4.56	12.40	11.37	10.34	20.45	0.00	14.50	9.68	10.32	46.78	23.60	16.67
AudioGPT-GPT4	22.76	25.05	36.89	36.48	43.10	32.18	39.78	37.09	51.10	25.34	25.70	29.80	25.30
SpeechGPT-7B-com	20.34	34.23	24.34	28.35	0.00	65.40	45.60	13.09	0.00	20.40	0.00	67.50	16.34
LLaMA-Omni	27.60	22.89	36.90	41.42	83.00	74.56	80.20	43.30	63.20	28.37	84.40	65.30	19.78
3D-LLM	4.24	1.03	17.38	8.69	48.20	33.13	40.00	12.15	51.20	6.40	35.60	30.00	5.32
PointLLM-7B	16.65	23.89	16.02	17.27	48.20	33.13	40.00	20.15	51.20	20.26	35.60	30.00	18.32
PointLLM-13B	17.24	23.20	16.45	16.99	45.80	32.93	32.20	20.01	51.20	19.91	34.80	30.60	18.39
3D-VisTA	0.20	0.00	0.31	0.05	48.20	33.13	40.00	0.23	51.20	0.10	35.60	30.00	0.05
MotionGPT-T5	1.27	0.00	5.00	0.91	48.20	33.13	40.00	1.82	51.20	1.13	35.60	30.00	0.55
MotionGPT-LLaMA	18.53	19.87	8.41	6.14	43.40	33.54	41.80	11.64	46.00	14.46	39.40	39.20	21.27
AvatarGPT	7.99	2.38	13.69	7.64	48.20	33.13	40.00	11.84	51.20	7.48	35.60	30.00	6.09
LLaMA-Mesh	28.67	22.44	23.21	34.27	71.60	76.16	73.00	27.88	69.60	29.28	90.80	86.40	37.13

Table 75: Results on NLP Group, from #L-5 to #L-7.

Model	#L-5 (Non-Trad QA)			#L-6 (Advance QA)					#L-7 (Math Ability)		
	#1↑	#2↑	#3↑	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#2↑	#3↑
SoTA Specialist	43.46	77.01	53.59	48.87	61.77	83.33	32.01	88.00	75.00	81.00	69.34
Meta-Llama-3.1-8B-Instruct	26.80	18.10	15.14	17.70	17.43	46.32	33.22	71.20	48.60	29.20	30.89
Qwen2.5-7B-Instruct	28.44	17.60	14.79	18.52	20.58	18.50	18.91	70.60	32.80	11.20	26.82
Gemma-2-9b-it	33.64	21.78	15.99	20.32	21.18	57.76	29.47	72.60	46.20	24.20	34.41
ChatGLM-6b	29.10	21.31	7.75	26.56	24.08	12.04	21.72	54.80	24.40	0.00	32.15
Vicuna-7b-v1.5	3.30	6.10	4.14	2.27	0.73	1.05	2.19	60.00	23.40	0.00	36.43
InternLM-Chat-7b	30.64	22.06	9.71	22.50	26.23	12.30	21.58	71.60	26.20	1.80	33.92
GPT-J-6B	2.03	1.86	0.93	1.08	0.05	0.41	1.22	56.40	24.40	0.00	25.68
Falcon3-7B-Instruct	26.33	23.70	14.10	18.79	21.22	49.90	25.67	70.00	27.20	35.00	33.89
Baichuan2-7B-Base	1.12	1.11	4.19	0.52	0.69	9.47	0.61	63.00	24.40	22.40	29.18
Ministral-8B-Instruct-2410	34.01	42.06	41.87	36.71	45.25	15.73	29.36	75.40	30.80	0.00	8.19
Yi-Ligntning	25.42	20.37	15.79	16.91	19.55	55.99	20.86	85.83	25.80	18.60	22.95
GPT-3.5-turbo	30.46	12.26	5.86	18.50	12.41	27.83	20.27	64.00	24.40	0.00	43.85
GPT-4v	25.48	10.78	7.04	12.22	10.99	34.89	17.12	64.00	24.40	37.05	21.91
GPT-4o	29.30	11.67	7.64	13.55	11.79	22.52	19.45	64.00	24.40	35.05	21.97
GPT4o-mini	27.23	11.65	7.28	12.92	12.92	23.57	16.45	64.00	24.40	34.42	20.65
GPT-4o-4096	29.82	13.40	8.30	13.97	13.12	24.63	23.01	64.00	24.40	0.00	20.47
ChatGPT-4o-latest	30.07	13.75	7.05	14.22	12.06	28.23	22.35	64.00	24.40	0.00	18.59
Claude-3.5-Sonnet	31.26	40.07	28.78	32.27	35.84	32.42	39.32	64.60	41.40	52.00	47.62
Claude-3.5-Opus	25.90	41.43	26.53	29.89	28.32	26.08	32.91	59.20	39.00	39.60	42.33
Emu2-32B	26.88	34.62	21.72	22.77	25.92	25.93	28.18	53.40	37.80	40.60	38.71
DetGPT	19.99	31.33	17.94	20.13	19.03	20.62	27.87	50.00	34.60	32.00	29.10
InternVL2.5-8B	31.39	45.60	51.60	39.12	25.42	17.02	50.29	31.76	51.66	52.69	48.58
InternVL2.5-4B	42.78	69.20	44.60	38.83	26.18	8.28	50.89	31.73	50.05	31.95	41.91
NExT-GPT-V1.5	3.30	5.40	3.87	1.67	0.73	9.05	11.60	59.30	20.80	0.00	28.67
InternVL2.5-2B	26.67	60.20	36.80	33.95	22.26	15.69	50.69	28.50	38.05	46.41	38.50
Monkey-10B-chat	3.61	57.60	19.60	35.77	31.43	20.13	53.69	24.29	55.51	29.66	51.58
DeepSeek-VL-7B	21.28	70.60	29.60	38.93	38.93	70.60	70.60	70.60	70.60	57.66	21.28
Qwen2-VL-7B	28.90	15.17	9.45	14.99	18.27	3.21	9.33	59.00	28.00	41.96	38.80
Qwen-VL-Chat	2.77	2.71	2.83	1.71	0.61	0.14	1.70	63.70	22.60	7.78	27.42
Qwen-Audio-Chat	25.48	18.21	11.82	16.40	16.65	5.38	16.29	61.40	24.60	0.00	34.70
Qwen2-Audio-Instruct	25.10	20.82	9.67	17.08	18.48	14.83	22.94	72.30	0.00	0.00	36.75
MoE-LLAVA-Phi2-2.7B-4e-384	24.98	22.15	15.64	18.48	20.99	31.77	21.40	55.40	25.00	43.29	27.91
mPLUG-Owl2-LLaMA2-7b	28.29	24.61	11.98	22.22	31.37	8.96	11.52	53.72	27.00	36.59	30.15
Phi-3.5-Vision-Instruct	30.25	35.99	20.52	33.28	34.39	54.86	27.17	59.44	30.40	44.65	42.87
Cambrian-1-8B	22.34	15.39	10.86	15.86	16.65	28.93	14.32	48.28	24.40	42.66	34.28
MiniGPT4-LLaMA2	3.19	3.57	4.36	2.03	0.72	0.16	1.93	50.13	31.20	7.78	38.93
InternVL-Chat-V1-5	17.79	63.00	37.20	2.20	42.92	14.65	52.49	30.51	9.17	51.42	50.62
Mini-InternVL-Chat-4B-V1-5	32.81	63.80	23.20	5.00	44.55	22.60	50.09	30.43	16.63	1.30	44.39
InternLM-XComposer2-VL-1.8B	16.86	56.80	28.60	0.00	24.36	12.91	53.89	27.50	25.00	0.00	44.65
GPT4RoI	1.00	63.60	24.80	0.00	24.10	0.00	46.30	5.37	20.01	0.00	10.79
GLaMM	2.96	64.00	24.40	0.00	0.00	10.70	46.30	25.00	25.00	0.00	12.95
LLaVA-NeXT-13B	25.73	22.48	8.39	23.73	19.67	18.19	18.58	57.20	22.80	2.60	28.53
LLaVA-NeXT-34B	25.01	23.54	10.19	18.67	22.67	17.56	19.65	67.40	25.40	4.20	35.85
Pixtral-12B	27.67	27.40	11.43	23.07	20.02	24.11	22.53	63.80	27.20	4.80	36.67
SEED-LLaMA-13B	14.78	8.88	2.75	10.86	13.24	13.61	13.12	43.40	14.60	1.20	19.23
BLIP2	29.63	33.93	11.57	31.67	22.47	30.97	26.53	67.60	27.80	0.00	35.29
MiniMonkey	24.32	20.98	8.68	17.47	24.01	28.28	19.85	55.20	31.80	3.20	36.68
DeepSeek-VL-7B	9.17	64.40	27.00	0.40	39.87	7.19	47.70	24.55	46.70	21.73	17.69
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	25.04	37.77	24.74	25.41	26.30	23.17	33.51	57.40	41.60	42.80	39.87
ShareGPT4V-7B	23.19	28.22	15.83	18.39	16.80	19.34	23.24	48.40	36.00	35.20	32.22
ShareGPT4V-13B	23.23	29.06	21.96	26.99	24.93	23.99	29.35	51.60	35.60	40.60	38.75

Model	#L-5 (Non-Trad QA)			#L-6 (Advance QA)					#L-7 (Math Ability)		
	#1↑	#2↑	#3↑	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#2↑	#3↑
BLIP-3 (XGen-MM)	24.30	34.92	24.63	26.70	21.70	23.61	33.51	53.40	35.20	41.60	38.25
AnyGPT	10.04	13.99	0.00	11.03	7.64	8.71	13.25	18.80	14.20	13.00	13.45
MiniCPM3-4B	32.94	37.32	22.96	24.09	29.80	27.26	31.46	55.20	43.40	48.20	41.36
LaVIT-V2 (7B)	21.39	31.47	18.52	24.23	24.22	23.68	32.52	39.40	28.20	29.20	30.68
GLM-VL-Chat	32.60	38.71	19.90	21.72	26.52	24.85	29.63	54.20	41.00	46.80	41.84
Gemini-1.5-Pro	29.58	23.54	14.52	34.49	32.95	23.57	25.68	66.60	29.40	38.20	36.80
Gemini-1.5-Flash	31.04	19.82	11.75	31.64	28.39	21.68	26.45	61.60	26.80	34.20	35.40
OMG-LLaVA-InternLM20B	7.09	9.38	8.61	6.24	8.62	6.21	18.07	18.80	15.80	6.40	17.80
Idefics3-8B-Llama3	30.14	22.15	12.56	32.20	26.36	18.62	22.47	60.20	25.20	33.20	31.80
Yi-Vision-v2	23.42	15.03	15.12	14.57	11.36	22.98	18.79	64.00	24.40	33.49	20.04
Qwen2-VL-72B	29.19	18.23	11.01	30.03	17.87	32.39	18.81	67.21	27.30	43.00	31.16
Otter	41.37	1.81	2.26	2.34	1.78	0.23	3.41	60.03	23.60	7.84	20.83
Show-o	10.81	30.80	6.87	19.14	24.04	11.25	8.43	37.82	31.00	6.57	7.14
NExT-Chat	14.70	3.92	7.27	6.41	6.93	0.00	3.96	64.00	24.40	0.00	8.57
InternVL2-26B	25.60	22.40	15.20	21.60	29.90	39.10	25.20	70.60	29.40	12.80	43.50
Qwen2-VL-72B	31.70	38.10	41.20	41.70	40.40	68.80	25.10	73.40	30.10	47.80	35.10
DeepSeek-VL-2-small	25.90	21.60	12.00	17.30	18.30	25.90	17.60	72.00	26.60	4.40	29.20
DeepSeek-VL-2	25.60	18.60	12.30	17.10	18.60	33.80	20.10	73.80	29.40	15.20	28.40
LLaVA-One-Vision-7B	33.10	27.90	12.60	28.80	35.90	63.50	35.40	71.40	44.20	7.00	45.20
LLaVA-One-Vision-72B	34.40	22.60	32.60	35.80	43.20	75.80	53.20	77.00	52.20	8.60	44.90
Sa2VA-8B	28.60	33.70	45.60	26.80	37.80	52.70	33.80	74.40	51.80	15.80	34.20
Sa2VA-26B	28.70	23.60	24.10	21.30	24.10	59.20	30.20	73.00	55.20	17.40	35.20
CoLVA-2B	27.60	28.40	9.10	24.30	30.90	20.60	23.70	61.60	23.60	5.40	27.50
CoLVA-4B	30.60	22.00	12.20	21.20	26.10	29.40	32.30	72.60	47.80	14.00	34.80
Long-LLaVA	22.50	22.40	8.60	18.00	31.10	31.70	18.30	66.60	24.20	0.60	37.10
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	3.10	14.30	3.87	13.76	0.73	9.05	16.10	53.90	21.20	0.00	26.87
PandaGPT (13B)	2.30	4.70	3.19	1.94	0.73	0.94	2.07	56.96	5.63	0.00	8.75
AnyGPT	10.04	13.99	0.00	11.03	7.64	8.71	13.25	18.80	14.20	13.00	13.45
GAMA	21.54	15.39	7.79	18.91	13.97	0.31	1.16	68.30	0.00	0.00	7.03
Pengi	19.78	13.67	4.58	15.34	13.89	0.23	1.02	58.33	0.00	0.00	3.98
SALMONN-7B	9.98	5.37	5.39	2.83	3.56	0.06	2.41	62.52	20.00	0.00	14.51
SALMONN-13B	10.32	5.68	6.89	3.21	5.76	0.00	3.20	48.15	20.40	0.00	15.78
WavLLM	20.73	14.67	6.45	16.78	14.34	0.32	1.07	29.78	0.00	0.00	11.45
ImageBind-LLM	10.45	7.56	4.34	6.78	8.45	0.34	1.04	35.22	0.00	0.00	10.29
Unified-io-2-XXL	1.20	11.10	5.67	0.00	0.89	0.00	13.56	55.40	10.40	0.00	6.57
ModaVerse-7b-v0	2.10	4.50	2.57	1.54	1.56	0.30	1.34	54.36	5.06	0.00	8.34
AudioGPT-GPT4	24.58	10.70	7.04	12.22	10.23	34.67	15.67	63.59	24.30	37.01	21.89
SpeechGPT-7B-com	18.34	14.67	6.45	15.34	13.45	0.20	1.00	78.60	0.00	0.00	8.35
LLaMA-Omni	26.80	18.10	15.14	16.70	7.45	6.23	32.23	73.40	48.60	29.20	30.89
3D-LLM	18.51	35.77	16.70	44.61	43.83	0.00	30.84	63.80	24.40	0.00	12.09
PointLLM-7B	3.27	6.15	4.13	2.20	0.72	0.77	2.13	64.00	24.40	0.00	34.64
PointLLM-13B	3.30	6.08	4.29	2.25	0.73	0.00	2.12	55.40	23.80	0.00	31.16
3D-VisTA	0.51	0.62	3.47	0.61	0.33	0.26	0.11	64.00	24.40	0.00	0.00
MotionGPT-T5	0.81	0.00	0.85	0.09	0.16	0.00	0.07	64.00	24.40	0.00	0.03
MotionGPT-LLaMA	4.75	2.24	3.56	2.31	1.68	1.52	1.66	55.20	24.80	0.00	14.34
AvatarGPT	8.53	2.16	3.77	1.59	1.58	0.11	1.19	64.00	24.40	0.00	0.90
LLaMA-Mesh	3.24	6.05	5.98	2.10	0.72	0.93	1.92	68.60	24.60	0.00	39.22

Table 77: Results on NLP Group, from #L-8 to #L-11.

Model	#L-8 (Code Ability)					#L-9 (X-Lan&NMT)				#L-10 (Txt Sum)			#L-11 (Dialog Gen)
	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#2↑	#3↑	#4↑	#1↑	#2↑	#3↑	#1↑
SoTA Specialist	80.50	79.50	75.06	76.94	77.20	71.06	83.61	72.83	91.30	48.16	76.23	91.35	28.27
Meta-Llama-3.1-8B-Instruct	19.72	47.31	32.72	37.18	8.67	54.06	61.57	21.24	76.06	47.56	35.00	50.92	14.80
Qwen2.5-7B-Instruct	11.72	54.49	35.18	53.19	54.39	56.55	66.23	30.63	76.70	47.90	33.47	49.16	15.34
Gemma-2-9b-it	18.11	49.10	33.06	59.71	55.22	53.25	71.21	11.99	79.77	36.49	36.88	50.18	13.34
ChatGLM-6b	11.13	46.71	22.25	56.28	43.05	23.50	39.20	24.54	24.16	43.80	22.69	49.04	13.93
Vicuna-7b-v1.5	3.79	53.69	24.58	52.13	42.18	53.36	60.79	30.15	74.92	45.00	28.18	48.56	5.06
InternLM-Chat-7b	7.33	53.69	26.96	54.17	1.50	42.63	51.67	25.74	59.02	43.33	26.05	47.63	13.01
GPT-J-6B	2.21	53.19	17.64	44.19	8.13	8.48	17.59	0.98	13.83	22.76	10.61	27.38	3.02
Falcon3-7B-Instruct	12.70	53.29	35.64	47.73	61.20	54.26	59.21	30.75	78.92	46.71	32.30	47.20	15.56
Baichuan2-7B-Base	18.02	46.51	33.39	49.72	50.66	37.06	54.76	13.24	58.91	22.42	12.63	20.43	4.92
Ministral-8B-Instruct-2410	16.74	49.90	12.77	5.34	29.54	56.93	70.39	21.21	78.95	47.77	33.55	49.06	13.73
Yi-Ligntning	20.95	49.70	37.06	56.55	53.59	60.59	74.60	31.47	83.43	43.32	32.40	50.38	15.29
GPT-3.5-turbo	17.51	54.09	33.62	39.57	37.16	63.11	77.81	33.81	86.27	46.69	32.20	50.18	14.40
GPT-4v	27.81	48.70	35.28	35.05	33.51	63.12	79.38	32.70	86.25	43.08	32.87	50.39	13.96
GPT-4o	28.47	51.70	33.80	36.01	33.20	64.63	81.15	34.15	87.52	44.59	32.80	50.67	14.49
GPT4o-mini	29.30	51.30	36.24	37.52	32.79	64.87	79.53	34.15	87.63	44.44	33.42	50.97	13.51
GPT-4o-4096	29.30	47.90	34.00	36.37	33.50	64.24	80.61	33.27	87.42	43.00	32.65	51.12	15.49
ChatGPT-4o-latest	28.60	49.30	35.93	35.82	33.28	64.18	80.93	33.16	87.38	44.10	33.39	50.28	15.08
Claude-3.5-Sonnet	33.84	54.49	51.79	59.84	41.81	55.89	52.89	31.70	65.44	43.18	44.77	73.60	20.70
Claude-3.5-Opus	37.47	51.30	46.75	56.24	45.08	57.25	52.53	29.57	63.71	36.54	46.37	62.96	17.39
Emu2-32B	32.04	44.91	45.72	46.81	39.13	52.12	47.20	25.60	56.98	35.04	37.31	66.70	13.05
DetGPT	21.64	38.92	37.46	42.17	33.02	42.92	40.35	20.22	50.60	33.39	33.55	57.84	11.12
InternVL2.5-8B	55.64	27.19	62.17	39.17	30.91	44.99	13.88	33.23	31.30	19.89	44.79	89.00	9.07
InternVL2.5-4B	50.96	25.71	57.20	42.76	29.77	46.58	12.58	26.61	30.53	24.94	41.01	89.68	7.82
NExT-GPT-V1.5	2.60	50.34	19.56	50.25	40.61	51.90	56.63	29.68	67.75	41.50	25.40	46.30	5.06
InternVL2.5-2B	32.80	27.27	47.17	40.82	27.67	42.10	12.37	35.38	27.72	22.89	42.25	86.33	9.07
Monkey-10B-chat	61.05	26.19	72.42	45.67	32.96	45.92	7.21	36.08	33.61	8.00	62.68	91.14	11.00
DeepSeek-VL-7B	61.84	30.88	72.05	47.26	31.63	47.87	14.46	43.94	1.83	30.94	51.35	88.00	10.37
Qwen2-VL-7B	7.19	52.00	21.93	40.53	24.82	29.55	51.07	15.98	45.08	35.21	23.71	47.81	12.62
Qwen-VL-Chat	2.37	53.69	26.76	45.80	9.56	35.24	50.27	15.34	63.81	24.47	16.08	40.63	12.83
Qwen-Audio-Chat	8.70	50.89	19.75	42.35	48.19	53.76	61.65	29.92	76.89	44.99	26.21	48.40	10.47
Qwen2-Audio-Instruct	13.57	53.69	20.76	43.56	40.36	57.88	60.68	28.59	78.96	41.77	29.40	48.45	12.08
MoE-LLAVA-Phi2-2.7B-4e-384	8.13	53.69	23.32	53.18	53.79	55.75	62.18	11.40	75.50	46.05	33.73	48.80	11.16
mPLUG-Owl2-LLaMA2-7b	8.50	46.91	31.44	50.86	43.54	31.63	27.47	12.00	50.07	43.70	31.77	48.15	10.65
Phi-3.5-Vision-Instruct	18.92	53.09	33.90	58.48	54.41	55.85	49.05	13.62	74.88	47.24	31.66	49.57	12.74
Cambrian-1-8B	2.37	53.49	24.93	63.10	14.25	36.46	60.68	10.69	54.16	34.43	28.95	34.16	8.90
MiniGPT4-LLaMA2	2.51	53.69	35.05	55.30	68.62	35.30	46.96	11.57	47.05	37.60	22.71	48.11	10.42
InternVL-Chat-V1-5	57.70	30.73	65.14	44.44	44.44	47.94	13.49	43.81	28.73	16.33	39.75	83.44	0.00
Mini-InternVL-Chat-4B-V1-5	41.39	24.82	61.32	43.74	29.68	46.01	12.77	36.12	29.01	13.13	43.70	84.72	0.00
InternLM-XComposer2-VL-1.8B	38.12	26.67	58.53	46.80	32.32	46.73	12.45	18.42	18.91	22.65	36.37	89.48	0.00
GPT4RoI	15.18	7.10	17.79	22.13	13.56	20.58	4.23	9.92	17.10	2.52	17.63	23.03	0.00
GLaMM	20.84	0.00	5.40	16.10	13.26	21.56	7.96	2.14	4.94	0.00	21.51	40.08	0.00
LLaVA-NeXT-13B	6.32	36.33	20.27	33.54	19.56	36.23	32.17	8.47	36.07	32.18	25.54	44.36	9.57
LLaVA-NeXT-34B	8.54	34.93	29.87	45.67	27.05	40.98	35.86	14.76	57.21	36.32	34.89	45.60	12.88
Pixtral-12B	11.12	48.70	27.43	40.28	26.79	37.43	36.27	13.24	45.04	36.25	32.56	44.09	12.00
SEED-LLaMA-13B	6.32	27.94	15.79	22.01	7.39	26.19	32.94	7.26	20.15	24.31	14.23	32.14	4.80
BLIP2	3.67	22.18	16.73	29.64	6.22	37.23	31.77	13.81	48.51	33.50	34.77	44.32	8.56
MiniMonkey	14.91	53.09	28.92	51.37	49.53	46.17	39.55	31.19	63.30	42.16	29.16	49.10	12.73
DeepSeek-VL-7B	34.22	4.67	23.61	29.68	16.64	43.10	11.22	30.94	15.02	20.13	17.55	67.17	11.07
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	27.55	44.31	46.48	46.92	40.04	52.95	50.06	26.57	58.20	35.14	41.01	61.89	16.03
ShareGPT4V-7B	25.34	40.32	39.39	44.56	33.93	43.31	38.89	20.14	49.65	33.37	32.77	57.34	10.51
ShareGPT4V-13B	27.13	45.11	40.78	46.31	34.36	48.81	43.03	23.19	55.32	36.16	34.16	61.82	12.86

Model	#L-8 (Code Ability)					#L-9 (X-Lan&NMT)				#L-10 (Txt Sum)			#L-11 (Dialog Gen)
	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#2↑	#3↑	#4↑	#1↑	#2↑	#3↑	#1↑
BLIP-3 (XGen-MM)	29.16	44.71	46.65	50.89	39.95	50.52	44.40	23.96	57.25	34.78	34.46	63.27	13.21
AnyGPT	12.37	17.40	17.53	17.56	14.45	18.44	17.13	9.47	19.59	12.22	15.13	21.36	0.00
MiniCPM3-4B	35.26	43.60	48.38	47.14	45.70	46.22	45.74	34.05	54.93	37.62	41.81	65.77	19.79
LaVIT-V2 (7B)	24.75	39.80	36.70	45.34	28.72	39.30	33.25	26.56	41.02	28.57	35.75	54.39	13.20
GLM-VL-Chat	30.68	44.40	48.68	48.28	45.02	47.93	45.57	31.90	53.90	34.14	38.47	63.65	20.85
Gemini-1.5-Pro	25.60	52.30	35.24	40.91	43.54	58.65	68.28	38.72	85.46	44.08	32.73	49.32	14.27
Gemini-1.5-Flash	23.21	48.70	33.05	36.72	33.20	52.75	64.15	32.93	80.12	38.24	31.34	50.15	12.74
OMG-LLaVA-InternLM20B	9.38	14.98	25.23	26.27	18.30	25.69	9.41	7.47	23.35	12.62	12.47	26.74	3.02
Idefics3-8B-Llama3	18.49	51.30	32.80	36.87	32.96	47.93	52.87	14.33	68.93	36.85	28.12	51.32	11.08
Yi-Vision-v2	5.47	53.09	31.72	35.39	31.72	60.47	72.07	21.26	54.83	46.97	27.92	48.72	12.69
Qwen2-VL-72B	25.07	52.66	32.31	39.28	50.83	64.91	82.45	28.42	84.52	45.12	31.45	49.72	17.27
Otter	5.43	47.33	17.20	18.67	9.76	9.96	30.37	2.25	9.22	28.02	17.26	24.31	6.51
Show-o	3.41	23.81	3.65	7.61	1.37	4.43	13.22	5.89	9.60	10.73	0.00	0.00	0.00
NExT-Chat	0.00	46.30	4.54	25.11	2.39	6.91	12.16	3.62	10.40	27.04	20.52	29.58	8.18
InternVL2-26B	18.20	51.10	31.80	50.80	54.10	41.30	54.50	26.90	61.00	44.20	29.70	48.20	13.80
Qwen2-VL-72B	24.80	53.40	32.30	52.10	26.10	61.10	75.40	30.60	84.10	45.30	31.70	49.90	17.80
DeepSeek-VL-2-small	13.80	46.90	29.80	66.20	49.40	56.00	60.70	30.80	73.60	41.30	29.90	46.00	15.20
DeepSeek-VL-2	20.00	44.70	31.90	62.60	55.50	56.10	64.60	31.30	75.30	43.50	30.10	46.80	16.30
LLaVA-One-Vision-7B	21.50	53.70	34.10	63.10	59.50	53.30	62.40	29.50	75.70	44.90	23.20	49.30	17.70
LLaVA-One-Vision-72B	11.80	53.50	35.30	45.30	70.50	56.40	70.60	18.30	75.70	44.10	29.40	50.90	17.70
Sa2VA-8B	14.40	52.30	35.60	59.90	56.60	57.30	62.80	31.10	76.00	45.60	33.20	49.60	18.60
Sa2VA-26B	13.40	51.30	35.60	58.50	45.50	59.70	60.90	34.20	79.80	25.70	33.30	50.50	16.90
CoLVA-2B	8.80	54.10	29.10	42.80	51.80	42.20	36.20	26.40	58.20	41.90	27.90	46.40	12.80
CoLVA-4B	11.80	53.70	34.00	52.10	55.10	54.30	60.50	29.70	74.50	44.70	30.90	49.40	17.20
Long-LLaVA	16.90	46.30	25.20	62.30	41.50	44.30	46.00	30.30	75.00	43.30	21.90	49.10	6.30
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	12.30	50.34	15.62	50.23	41.09	50.37	46.63	26.78	67.75	41.30	24.60	44.60	5.30
PandaGPT (13B)	0.68	45.37	24.58	52.13	54.58	7.56	5.37	3.84	22.39	7.37	3.89	6.46	0.00
AnyGPT	12.37	17.40	17.53	17.56	14.45	18.44	17.13	9.47	19.59	12.22	15.13	21.36	0.00
GAMA	5.49	53.69	18.79	34.56	67.50	0.08	0.59	0.00	0.03	0.38	0.34	0.65	7.01
Pengi	3.54	23.56	14.34	24.78	67.50	0.05	0.34	0.00	0.05	0.23	0.27	0.04	4.58
SALMONN-7B	1.11	53.89	10.89	25.56	56.36	11.43	12.42	6.40	15.42	9.99	2.65	9.87	1.15
SALMONN-13B	1.39	55.89	11.24	27.34	48.33	12.54	13.69	6.56	17.65	11.54	2.99	11.34	1.37
WavLLM	5.60	54.67	19.20	30.68	67.50	0.09	0.56	0.00	0.06	0.36	0.45	0.58	6.45
ImageBind-LLM	5.46	46.78	15.45	29.58	67.50	0.06	0.34	0.00	0.03	0.00	0.00	0.00	6.32
Unified-io-2-XXL	3.67	6.70	10.46	15.84	10.31	15.78	4.21	5.75	14.67	2.52	13.56	21.50	0.00
ModaVerse-7b-v0	0.68	43.50	23.60	46.60	54.58	7.56	5.37	3.84	22.39	7.37	3.89	6.46	0.00
AudioGPT-GPT4	27.56	48.50	35.46	35.45	33.51	63.05	78.50	33.24	84.34	43.08	32.45	51.23	14.05
SpeechGPT-7B-com	5.67	45.67	20.34	36.59	67.50	0.05	0.61	0.00	0.03	0.45	0.41	0.23	6.79
LLaMA-Omni	19.72	47.31	32.72	37.18	8.67	54.06	61.57	21.24	16.45	17.56	15.00	20.94	4.80
3D-LLM	2.13	46.31	0.00	0.00	0.65	27.70	35.62	51.18	4.04	32.50	29.52	22.22	4.87
PointLLM-7B	3.69	53.69	0.00	0.00	31.93	30.18	50.53	20.17	63.64	42.48	25.13	40.92	4.79
PointLLM-13B	3.53	53.69	0.00	0.00	41.05	42.57	55.99	25.94	70.58	44.87	30.72	41.31	4.71
3D-VisTA	0.65	46.31	0.00	0.00	0.01	0.86	2.93	0.06	0.05	2.41	2.42	1.70	0.21
MotionGPT-T5	0.23	46.31	4.94	0.05	0.40	0.02	0.01	0.08	0.02	0.31	0.66	0.46	0.26
MotionGPT-LLaMA	2.58	46.91	20.50	0.34	5.88	18.54	26.14	45.71	7.41	41.80	20.25	43.45	3.50
AvatarGPT	10.73	46.31	3.63	1.06	0.98	2.24	8.00	0.00	1.64	9.50	9.42	8.44	8.63
LLaMA-Mesh	3.86	53.69	30.35	53.56	54.75	54.69	72.44	26.26	76.21	45.89	28.61	47.44	4.93

Table 79: Results on NLP Group, from #L-12 to #L-16.

Model	#L-12 (TxT Gen)					#L-13 (Time Series)	#L-14 (Txt Cls)	#L-15 (Txt Entail)	#L-16 (Sem Analy)			
	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#1↑	#1↑	#1↑	#2↑	#3↑	#4↑
SoTA Specialist	93.02	75.19	83.46	85.89	97.19	0.31	94.40	91.41	88.00	98.00	85.40	72.80
Meta-Llama-3.1-8B-Instruct	39.83	29.90	31.12	42.10	83.74	7.95	76.40	51.80	44.79	80.00	75.00	63.80
Qwen2.5-7B-Instruct	50.67	33.58	31.85	47.24	90.90	5.74	88.20	79.35	66.40	49.00	63.80	46.80
Gemma-2-9b-it	36.37	25.40	22.22	9.33	86.00	5.75	84.46	55.82	75.60	83.80	62.20	48.00
ChatGLM-6b	32.84	27.89	30.32	48.30	74.85	10.91	41.80	45.81	50.20	47.80	0.00	0.00
Vicuna-7b-v1.5	30.01	30.10	23.34	46.16	90.29	11.41	0.00	0.00	0.00	0.00	0.00	0.00
InternLM-Chat-7b	32.59	26.43	22.48	9.72	91.15	10.23	77.80	6.87	58.80	14.20	31.00	3.60
GPT-J-6B	13.98	9.38	31.53	10.69	9.46	11.41	0.00	0.00	0.00	0.00	0.00	0.00
Falcon3-7B-Instruct	44.34	35.09	26.53	44.22	90.59	5.15	88.80	85.89	64.20	26.80	44.40	47.20
Baichuan2-7B-Base	49.50	36.31	14.53	55.40	93.31	7.85	0.00	0.00	0.00	0.00	0.00	0.00
Ministral-8B-Instruct-2410	15.54	16.33	32.30	23.68	29.11	11.08	84.80	72.60	53.80	67.00	61.60	44.40
Yi-Ligntning	52.74	34.45	28.52	52.30	95.39	5.37	72.60	56.24	78.20	70.60	66.40	43.80
GPT-3.5-turbo	35.00	31.90	27.20	38.32	83.48	1.55	73.80	63.19	50.20	41.80	55.80	28.00
GPT-4v	40.36	32.15	28.03	39.08	83.20	3.16	86.20	83.23	85.60	80.60	66.20	28.00
GPT-4o	43.98	33.99	29.40	41.49	83.20	2.58	85.40	86.30	71.40	90.20	69.00	39.40
GPT4o-mini	42.97	32.93	25.60	42.10	82.31	4.24	82.00	86.50	82.00	84.40	58.60	47.40
GPT-4o-4096	43.32	34.24	28.16	43.04	83.06	3.08	87.40	83.84	80.20	88.00	71.20	40.40
ChatGPT-4o-latest	41.21	32.09	22.33	42.67	82.63	2.73	88.60	77.51	80.40	89.60	69.00	41.60
Claude-3.5-Sonnet	59.94	48.94	44.02	47.91	90.18	3.88	68.59	52.58	47.87	51.86	59.26	41.41
Claude-3.5-Opus	58.00	47.16	35.40	42.21	88.65	3.67	30.74	30.74	30.74	30.74	30.74	30.74
Emu2-32B	51.27	43.36	31.58	42.43	82.09	9.53	57.54	48.78	44.02	46.12	50.11	34.79
DetGPT	44.80	38.08	26.30	37.63	76.72	16.87	52.97	46.25	42.05	43.83	44.60	30.15
InternVL2.5-8B	87.00	72.20	54.60	54.60	91.40	75.20	55.40	68.40	48.80	52.40	82.80	42.99
InternVL2.5-4B	81.80	70.40	69.00	69.20	90.20	70.80	46.00	61.80	48.20	54.60	80.60	39.20
NExT-GPT-V1.5	25.40	30.40	22.98	45.23	86.43	1.06	68.90	43.20	32.43	28.90	33.65	20.15
InternVL2.5-2B	83.20	69.20	25.40	30.20	85.40	79.60	50.60	51.40	48.20	51.60	83.60	42.40
Monkey-10B-chat	38.60	8.80	0.00	0.00	65.60	20.40	0.20	3.00	10.60	4.00	16.80	8.60
DeepSeek-VL-7B	78.40	68.80	83.00	83.00	85.20	83.00	62.20	50.60	54.20	54.20	84.40	56.40
Qwen2-VL-7B	32.99	17.98	14.83	34.13	86.23	6.48	64.00	37.00	0.00	0.00	2.00	12.00
Qwen-VL-Chat	14.75	22.33	32.49	8.70	8.24	7.46	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-Audio-Chat	44.56	34.02	30.28	45.78	89.65	0.98	74.50	58.43	59.30	37.50	50.10	45.30
Qwen2-Audio-Instruct	43.34	32.16	30.73	40.75	88.39	1.23	77.60	58.69	61.00	38.40	51.40	47.40
MoE-LLAVA-Phi2-2.7B-4e-384	38.88	31.12	2.10	40.69	91.02	6.35	64.20	60.94	58.80	78.20	53.00	43.00
mPLUG-Owl2-LLaMA2-7b	34.43	33.43	28.22	44.03	90.75	6.16	76.20	38.04	36.20	25.80	59.00	32.40
Phi-3.5-Vision-Instruct	47.43	35.54	27.99	42.16	92.78	3.80	58.40	68.92	70.40	77.00	59.20	46.80
Cambrian-1-8B	31.08	27.58	2.68	41.38	90.04	3.14	59.80	0.00	13.40	30.40	19.80	3.20
MiniGPT4-LLaMA2	28.55	34.40	32.58	31.91	85.35	7.46	0.00	0.00	0.00	0.00	0.00	0.00
InternVL-Chat-V1-5	65.80	73.20	69.80	68.40	89.20	73.80	48.40	65.80	54.40	55.60	79.80	30.00
Mini-InternVL-Chat-4B-V1-5	65.80	44.60	53.40	56.80	84.80	68.60	48.60	50.80	50.20	52.40	64.40	72.40
InternLM-XComposer2-VL-1.8B	72.80	60.60	52.40	52.40	85.60	5.60	0.00	3.20	0.00	0.00	0.00	21.20
GPT4RoI	0.00	0.00	0.00	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	30.70	21.54	23.34	28.77	70.14	10.24	77.24	68.83	49.40	66.80	48.80	45.00
LLaVA-NeXT-34B	27.31	25.17	21.52	30.19	79.86	8.85	74.83	72.25	49.60	72.60	53.40	46.20
Pixtral-12B	34.42	26.08	23.77	33.09	77.35	8.12	80.91	66.31	54.80	71.20	51.80	44.80
SEED-LLaMA-13B	17.58	13.29	13.04	17.92	42.39	11.16	13.20	34.80	30.20	32.20	25.10	28.40
BLIP2	27.41	22.16	25.04	18.99	84.01	6.43	79.00	66.20	48.40	73.80	52.20	46.40
MiniMonkey	37.57	33.30	21.51	52.99	88.93	10.57	84.00	68.92	46.00	74.60	50.80	46.80
DeepSeek-VL-7B	1.40	0.00	0.80	2.80	0.35	∞	0.80	28.40	32.00	2.20	0.00	0.00
LISA-7B	0.00	0.00	0.00	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	53.67	42.22	31.42	42.56	81.70	20.45	53.55	51.28	45.25	47.31	49.32	29.67
ShareGPT4V-7B	48.38	36.65	30.58	38.87	79.03	27.95	49.45	44.62	37.27	44.93	48.38	27.06
ShareGPT4V-13B	52.42	39.82	29.59	38.93	81.80	24.69	54.03	49.62	45.91	43.94	48.08	34.05

Model	#L-12 (TxT Gen)					#L-13 (Time Series)	#L-14 (Txt Cls)	#L-15 (Txt Entail)	#L-16 (Sem Analy)			
	#1↑	#2↑	#3↑	#4↑	#5↑	#1↑	#1↑	#1↑	#1↑	#2↑	#3↑	#4↑
BLIP-3 (XGen-MM)	51.43	38.84	31.86	41.39	84.73	21.54	55.67	48.61	47.80	48.81	53.39	31.05
AnyGPT	17.27	16.66	12.12	15.78	28.03	33.39	19.48	16.48	17.92	18.21	17.06	11.55
MiniCPM3-4B	52.48	51.50	43.03	46.59	91.38	19.62	57.84	52.66	46.27	55.05	54.65	36.88
LaVIT-V2 (7B)	45.79	36.18	31.67	40.84	67.22	24.13	42.71	37.63	33.18	38.75	43.22	26.92
GLM-VL-Chat	54.43	49.68	41.76	44.94	88.66	21.37	58.65	49.30	46.31	52.65	53.46	37.94
Gemini-1.5-Pro	52.98	41.15	37.65	46.32	92.34	4.78	84.20	83.70	82.40	86.40	64.80	47.20
Gemini-1.5-Flash	49.43	33.48	30.06	41.47	88.95	7.92	80.60	77.41	74.20	80.20	58.60	39.80
OMG-LLaVA-InternLM20B	24.83	21.02	15.89	18.73	52.60	56.52	26.20	19.08	28.60	20.40	21.60	13.20
Idefics3-8B-Llama3	47.65	34.92	29.01	42.15	84.39	4.31	74.20	76.90	71.00	78.20	60.40	38.40
Yi-Vision-v2	35.39	29.41	25.08	36.80	76.91	7.32	59.60	49.69	63.20	85.40	60.20	42.60
Qwen2-VL-72B	31.34	31.24	37.84	38.85	84.19	4.45	90.00	80.67	85.33	82.00	69.33	44.00
Otter	13.98	17.65	9.29	11.66	14.88	11.51	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	11.20	4.53	3.11	8.26	4.35	∞	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	11.43	16.88	1.92	9.43	26.49	11.54	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2-26B	41.20	31.20	15.40	38.90	88.40	64.60	80.20	79.90	64.00	64.80	52.00	42.40
Qwen2-VL-72B	42.50	35.30	0.30	42.30	94.70	3.90	83.60	83.60	84.60	86.60	68.00	43.20
DeepSeek-VL-2-small	39.30	34.60	28.20	44.90	92.60	12.90	60.40	55.00	59.00	86.20	61.80	46.40
DeepSeek-VL-2	41.00	35.90	26.20	44.50	94.00	8.80	81.80	68.50	64.20	85.60	60.80	49.40
LLaVA-One-Vision-7B	29.90	34.30	11.20	42.80	93.90	5.90	77.40	65.40	65.20	60.40	45.80	47.80
LLaVA-One-Vision-72B	51.80	32.30	11.30	39.70	83.90	3.60	84.80	10.40	82.40	61.20	64.40	29.40
Sa2VA-8B	41.40	35.90	26.00	45.40	93.80	10.20	86.80	76.90	83.80	65.40	59.00	47.20
Sa2VA-26B	41.50	36.10	27.20	46.10	94.40	5.70	86.60	85.10	75.20	59.00	65.00	44.60
CoLVA-2B	30.90	33.10	21.20	54.90	90.00	11.00	75.40	3.10	30.40	67.30	25.40	35.80
CoLVA-4B	47.60	33.70	24.90	44.80	92.30	9.80	84.40	76.90	79.00	66.20	27.60	45.60
Long-LLaVA	41.20	29.80	31.60	50.40	89.20	11.40	68.60	41.70	41.40	66.80	56.20	46.20
LM4LV	0.00	0.00	0.00	0.00	0.00	∞	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	27.80	30.20	23.10	45.68	85.93	2.47	68.70	43.20	33.47	26.80	31.51	23.62
PandaGPT (13B)	4.78	5.47	6.07	36.48	28.67	0.41	0.00	40.80	0.00	0.00	15.60	0.00
AnyGPT	17.27	16.66	12.12	15.78	28.03	33.39	19.48	17.92	18.21	17.06	11.55	19.32
GAMA	8.11	0.29	1.92	6.04	7.74	0.78	0.00	0.00	0.00	0.00	0.00	0.00
Pengi	7.94	0.05	0.89	10.89	6.43	0.45	0.00	0.00	0.00	0.00	0.00	0.00
SALMONN-7B	4.38	9.57	6.03	30.93	81.93	0.78	1.40	0.61	6.60	0.00	22.20	7.20
SALMONN-13B	4.86	10.76	6.45	30.78	84.67	1.56	0.00	0.00	0.00	0.00	15.60	0.00
WavLLM	18.95	0.46	1.64	11.32	5.46	0.69	0.00	0.00	0.00	0.00	5.40	0.00
ImageBind-LLM	13.27	0.21	1.23	5.78	3.26	0.83	0.00	0.00	0.00	0.00	3.40	0.00
Unified-io-2-XXL	0.00	0.00	0.00	0.00	0.00	∞	23.50	34.60	44.30	5.60	23.50	0.00
ModaVerse-7b-v0	4.78	5.47	6.07	36.48	28.67	0.41	0.00	40.80	0.00	0.00	15.60	0.00
AudioGPT-GPT4	40.23	33.16	28.04	39.05	84.34	3.15	85.90	84.53	84.70	86.50	66.40	29.70
SpeechGPT-7B-com	8.94	0.37	1.56	15.78	9.29	0.58	0.00	0.00	0.00	0.00	0.00	0.00
LLaMA-Omni	3.89	9.90	31.12	42.10	68.40	5.34	37.50	35.80	43.81	60.30	25.40	34.50
3D-LLM	2.08	24.46	2.85	59.26	65.27	∞	47.00	33.95	0.40	0.00	51.00	53.60
PointLLM-7B	27.05	27.77	1.32	31.27	79.90	∞	0.00	0.00	0.00	0.00	0.00	0.00
PointLLM-13B	28.84	28.58	1.01	35.02	78.19	∞	0.00	0.00	0.00	0.00	0.00	0.00
3D-VisTA	0.00	0.30	0.00	1.31	1.45	∞	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-T5	0.00	0.08	0.00	0.86	5.92	11.11	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-LLaMA	19.65	7.15	20.58	14.94	9.31	11.06	0.00	0.00	0.00	0.00	0.00	0.00
AvatarGPT	1.65	12.21	0.09	8.15	7.88	11.11	0.00	0.00	0.00	0.00	0.00	0.00
LLaMA-Mesh	30.73	32.20	31.62	41.21	85.18	11.41	0.00	0.00	0.00	0.00	0.00	0.00

Table 81: Results on NLP Group, #L-17, part A.

Model	#L-17 (Affect Computing)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
SoTA Specialist	93.40	95.80	97.33	95.80	96.80	97.40	85.93	95.80	84.71
Meta-Llama-3.1-8B-Instruct	60.40	53.40	79.60	83.60	75.20	51.80	47.60	50.66	80.20
Qwen2.5-7B-Instruct	53.20	63.60	92.40	79.80	77.40	57.40	27.31	86.40	22.73
Gemma-2-9b-it	56.80	68.60	89.40	77.80	74.60	55.40	58.92	88.40	41.41
ChatGLM-6b	3.00	0.00	83.80	0.00	60.20	0.00	30.70	22.20	19.86
Vicuna-7b-v1.5	0.00	0.00	0.00	0.00	0.00	0.00	16.80	0.00	0.30
InternLM-Chat-7b	49.00	59.40	59.00	19.60	63.00	1.80	30.76	49.00	17.25
GPT-J-6B	0.00	0.00	0.00	0.00	0.00	0.00	13.97	0.00	0.36
Falcon3-7B-Instruct	53.60	65.80	90.60	87.00	72.60	60.00	43.21	85.20	23.67
Baichuan2-7B-Base	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.18
Ministral-8B-Instruct-2410	53.00	60.20	83.20	73.00	60.00	52.40	42.80	84.40	19.50
Yi-Ligntning	46.60	73.60	94.20	85.20	78.20	59.00	48.38	88.40	14.00
GPT-3.5-turbo	53.80	65.80	92.60	83.40	69.60	52.20	46.65	89.60	41.10
GPT-4v	62.00	68.40	94.00	92.00	76.60	59.40	57.19	88.40	53.92
GPT-4o	59.60	80.20	92.20	92.00	80.00	60.40	56.09	85.60	58.60
GPT4o-mini	53.60	72.60	91.60	81.40	75.00	56.80	50.47	87.40	50.45
GPT-4o-4096	58.60	80.60	85.20	90.60	79.20	60.40	54.04	85.20	40.77
ChatGPT-4o-latest	57.40	80.20	92.20	91.00	78.00	59.40	57.20	87.00	55.85
Claude-3.5-Sonnet	65.61	51.42	67.78	72.64	60.45	51.51	54.36	71.67	39.57
Claude-3.5-Opus	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74
Emu2-32B	53.44	47.58	57.62	63.95	53.24	39.91	42.49	60.45	25.76
DetGPT	48.97	42.55	48.73	58.75	45.39	37.59	40.39	60.73	25.05
InternVL2.5-8B	92.00	45.70	55.60	75.25	78.60	41.99	56.25	86.80	35.21
InternVL2.5-4B	92.60	42.99	41.60	82.41	67.20	43.79	53.89	87.59	30.10
NExT-GPT-V1.5	0.00	15.34	0.00	21.34	18.59	13.29	15.78	0.00	0.30
InternVL2.5-2B	44.80	34.40	43.20	85.27	39.00	43.20	40.14	81.00	24.36
Monkey-10B-chat	45.40	25.40	5.40	37.21	49.60	0.60	39.34	0.60	6.73
DeepSeek-VL-7B	87.00	46.00	66.80	59.91	48.60	46.60	51.86	86.40	24.87
Qwen2-VL-7B	56.00	53.00	27.00	59.00	9.00	41.00	25.36	27.00	6.55
Qwen-VL-Chat	0.00	0.00	0.00	0.00	0.00	0.00	14.82	0.00	8.01
Qwen-Audio-Chat	46.60	45.80	68.30	47.50	74.50	31.90	42.50	83.40	19.01
Qwen2-Audio-Instruct	48.20	48.60	70.00	47.80	73.20	32.20	40.89	84.20	19.89
MoE-LLAVA-Phi2-2.7B-4e-384	49.80	56.00	92.20	80.00	92.20	47.40	40.25	92.20	21.41
mPLUG-Owl2-LLaMA2-7b	53.80	55.60	84.80	54.80	68.80	43.40	31.38	84.80	7.98
Phi-3.5-Vision-Instruct	57.60	50.60	88.60	86.20	60.20	58.20	59.53	88.60	24.95
Cambrian-1-8B	45.00	42.40	0.00	63.40	3.00	53.60	22.87	0.00	23.69
MiniGPT4-LLaMA2	0.00	0.00	0.00	0.00	0.00	0.00	18.67	0.00	8.98
InternVL-Chat-V1-5	82.60	39.80	61.00	77.70	66.20	38.80	52.50	65.40	38.67
Mini-InternVL-Chat-4B-V1-5	69.00	47.00	54.00	62.78	61.80	39.00	44.62	81.80	27.36
InternLM-XComposer2-VL-1.8B	0.00	40.20	20.40	75.66	24.20	1.60	28.34	67.20	24.27
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	0.00	0.00	0.00	32.31	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	47.20	51.60	81.60	60.20	62.20	53.40	34.65	82.80	27.98
LLaVA-NeXT-34B	56.80	52.20	82.20	63.40	65.50	56.80	37.59	83.20	31.87
Pixtral-12B	53.40	57.20	88.80	67.40	67.30	57.80	42.07	85.40	37.98
SEED-LLaMA-13B	27.80	35.40	57.40	16.75	39.60	26.20	22.53	44.60	17.43
BLIP2	51.60	53.60	84.40	64.80	67.20	51.80	29.43	83.00	18.33
MiniMonkey	57.80	54.80	90.60	61.60	64.60	54.40	37.92	83.80	22.22
DeepSeek-VL-7B	0.00	0.00	13.60	0.00	0.00	0.00	17.03	0.00	3.60
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	59.80	45.53	57.04	68.49	52.96	44.74	47.84	61.99	32.61
ShareGPT4V-7B	51.16	39.94	51.54	62.69	45.67	40.59	36.31	55.02	25.49
ShareGPT4V-13B	52.77	46.04	50.41	65.68	52.99	37.92	42.55	58.30	30.48

Model	#L-17 (Affect Computing)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
BLIP-3 (XGen-MM)	54.88	50.33	51.44	68.68	52.94	41.51	44.27	59.55	30.82
AnyGPT	19.32	0.00	20.85	22.88	20.28	16.73	17.69	23.17	10.67
MiniCPM3-4B	56.27	45.12	53.29	62.79	50.98	43.43	47.19	59.28	31.01
LaVIT-V2 (7B)	48.87	34.25	41.14	50.13	41.12	33.80	34.92	51.52	23.71
GLM-VL-Chat	56.44	45.58	55.63	60.20	53.13	42.98	49.83	57.31	27.52
Gemini-1.5-Pro	59.00	78.40	92.40	89.20	83.40	57.60	56.09	88.20	51.21
Gemini-1.5-Flash	53.40	74.60	88.20	85.60	75.60	52.20	52.11	84.20	38.73
OMG-LLaVA-InternLM20B	22.40	20.60	26.40	28.20	49.60	24.40	38.46	22.60	13.01
Idefics3-8B-Llama3	53.60	68.60	89.00	83.40	76.60	51.20	53.65	82.00	39.89
Yi-Vision-v2	50.00	64.40	89.00	82.00	48.40	54.60	47.46	87.40	30.27
Qwen2-VL-72B	59.33	78.66	96.00	90.00	79.33	56.67	67.40	82.00	55.80
Otter	0.00	0.00	0.00	0.00	0.00	0.00	18.40	0.00	0.80
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00	0.00
InternVL2-26B	58.60	61.80	90.20	80.60	66.10	54.20	38.90	76.40	44.00
Qwen2-VL-72B	55.40	77.60	94.40	90.80	77.60	58.40	59.80	85.20	57.10
DeepSeek-VL-2-small	51.20	49.00	91.00	49.00	52.80	44.40	35.60	88.60	18.60
DeepSeek-VL-2	54.40	49.80	88.80	77.00	66.60	50.20	41.10	88.20	31.70
LLaVA-One-Vision-7B	49.60	49.20	86.60	60.60	61.40	45.20	47.70	88.20	24.40
LLaVA-One-Vision-72B	33.20	22.40	91.20	49.40	77.80	0.00	57.10	19.80	38.90
Sa2VA-8B	57.80	65.80	92.00	79.20	74.00	56.00	52.70	82.00	34.50
Sa2VA-26B	52.40	77.40	93.60	86.60	78.80	57.40	40.00	86.20	14.00
CoLVA-2B	50.80	56.00	87.00	72.80	44.80	51.40	48.70	82.40	14.90
CoLVA-4B	50.60	56.00	90.00	56.00	76.60	51.60	36.40	83.60	25.70
Long-LLaVA	56.80	49.40	88.20	53.80	42.20	55.20	37.90	66.20	16.30
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	0.00	24.57	0.00	21.34	23.47	15.94	25.85	0.32	0.30
PandaGPT (13B)	5.70	0.40	0.00	0.00	2.60	0.80	0.00	0.00	0.10
AnyGPT	0.00	20.85	22.88	20.28	16.73	17.69	23.17	10.67	14.45
GAMA	3.70	0.40	0.00	0.00	0.00	0.80	15.78	0.00	0.30
Pengi	0.00	0.10	0.00	0.00	1.10	0.20	0.00	0.00	0.10
SALMONN-7B	10.00	3.80	1.20	0.80	3.20	2.20	21.42	3.00	12.13
SALMONN-13B	12.40	0.40	0.00	0.00	2.60	0.80	23.45	3.40	13.04
WavLLM	3.60	0.00	0.00	0.00	2.30	0.00	16.12	0.00	0.30
ImageBind-LLM	5.60	0.40	0.00	0.00	2.60	0.80	0.00	0.00	0.10
Unified-io-2-XXL	5.40	64.10	56.30	45.90	0.00	0.00	0.00	0.00	0.00
ModaVerse-7b-v0	5.70	0.40	0.00	0.00	2.60	0.80	0.00	0.00	0.10
AudioGPT-GPT4	65.90	69.50	93.60	89.60	75.40	53.51	56.14	86.49	54.37
SpeechGPT-7B-com	0.00	0.00	0.00	0.00	0.00	0.00	15.78	0.00	0.30
LLaMA-Omni	43.50	35.70	19.70	13.60	15.20	11.80	17.40	20.66	20.20
3D-LLM	0.40	0.20	84.20	17.40	55.60	0.00	2.95	52.80	16.49
PointLLM-7B	0.00	0.00	0.00	0.00	0.00	0.00	16.80	0.00	0.30
PointLLM-13B	0.00	0.00	0.00	0.00	0.00	0.00	16.82	0.00	0.30
3D-VisTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-T5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-LLaMA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.66	0.00
AvatarGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.00
LLaMA-Mesh	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.97	0.00

Table 83: Results on NLP Group, #L-17, part B.

Model	#L-17 (Affect Computing)								
	#10↑	#11↑	#12↑	#13↑	#14↑	#15↑	#16↑	#17↑	#18↑
SoTA Specialist	78.40	62.22	77.51	90.32	91.55	92.98	36.80	95.57	80.20
Meta-Llama-3.1-8B-Instruct	18.33	39.00	35.30	31.53	0.00	29.25	1.10	0.65	2.12
Qwen2.5-7B-Instruct	59.40	46.59	49.22	0.00	44.98	0.72	4.67	75.60	1.59
Gemma-2-9b-it	69.80	50.29	54.08	0.00	50.73	4.51	4.07	64.40	1.89
ChatGLM-6b	7.00	0.27	0.27	0.00	0.14	0.00	0.00	68.60	0.00
Vicuna-7b-v1.5	0.00	0.08	0.00	0.09	0.00	0.02	0.00	0.00	0.00
InternLM-Chat-7b	2.20	0.88	0.51	0.00	0.69	0.00	0.00	63.20	0.00
GPT-J-6B	0.00	0.06	0.00	0.07	0.03	0.02	0.00	0.00	0.00
Falcon3-7B-Instruct	24.60	31.51	31.91	0.00	30.03	0.00	3.02	68.20	0.46
Baichuan2-7B-Base	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ministral-8B-Instruct-2410	14.00	12.41	17.72	0.00	27.96	0.12	0.22	67.40	0.11
Yi-Ligntning	56.60	6.69	15.54	0.00	43.75	0.25	5.96	1.02	67.20
GPT-3.5-turbo	47.60	44.48	40.65	0.00	51.93	0.46	5.08	70.00	1.29
GPT-4v	67.80	42.53	50.08	0.00	68.14	0.90	12.11	73.00	2.35
GPT-4o	73.40	51.79	58.21	0.00	63.77	5.70	11.35	77.40	3.56
GPT4o-mini	64.20	52.58	57.97	0.00	61.64	2.32	9.14	72.60	1.40
GPT-4o-4096	72.80	52.57	58.51	0.00	69.79	6.23	10.55	74.00	3.51
ChatGPT-4o-latest	71.00	51.79	58.88	0.00	66.71	3.71	10.62	74.90	2.21
Claude-3.5-Sonnet	44.31	30.21	33.02	24.35	44.93	26.17	10.88	50.31	35.37
Claude-3.5-Opus	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74
Emu2-32B	35.85	19.48	25.50	11.72	40.02	14.10	5.42	39.99	23.62
DetGPT	31.89	14.01	18.71	9.20	33.16	9.58	2.75	34.90	18.84
InternVL2.5-8B	58.20	34.88	43.77	86.80	50.65	86.80	1.12	0.24	71.00
InternVL2.5-4B	48.40	38.29	35.97	0.00	33.47	1.64	0.73	0.42	72.60
NExT-GPT-V1.5	0.00	8.08	5.09	16.42	0.00	3.02	14.37	34.70	0.00
InternVL2.5-2B	37.40	15.14	12.96	0.00	10.30	0.20	0.00	0.11	72.40
Monkey-10B-chat	0.00	0.24	0.00	0.60	0.00	0.00	0.00	0.00	1.80
DeepSeek-VL-7B	40.80	3.95	5.12	86.40	36.00	86.40	0.15	0.15	66.60
Qwen2-VL-7B	12.00	0.00	0.00	0.00	26.00	0.00	0.00	27.00	0.00
Qwen-VL-Chat	0.00	0.19	0.24	0.06	0.00	0.02	0.00	0.00	0.00
Qwen-Audio-Chat	4.70	0.45	0.56	0.00	3.56	0.00	0.34	60.13	0.01
Qwen2-Audio-Instruct	4.80	0.58	0.68	0.00	4.16	0.00	0.26	60.60	0.09
MoE-LLAVA-Phi2-2.7B-4e-384	7.00	0.38	1.15	0.00	39.60	0.00	0.00	92.20	0.00
mPLUG-Owl2-LLaMA2-7b	37.20	1.82	0.25	0.00	76.40	0.00	0.00	84.80	0.00
Phi-3.5-Vision-Instruct	57.00	22.39	20.94	0.00	89.20	0.08	0.89	88.60	0.20
Cambrian-1-8B	6.20	13.05	4.13	0.00	0.20	0.14	0.91	0.00	0.18
MiniGPT4-LLaMA2	0.00	4.27	5.38	0.06	0.00	0.02	0.14	0.00	0.00
InternVL-Chat-V1-5	25.40	27.88	32.06	0.00	36.49	0.00	0.00	0.18	69.00
Mini-InternVL-Chat-4B-V1-5	41.80	7.72	16.57	0.00	29.42	0.00	0.00	0.00	61.00
InternLM-XComposer2-VL-1.8B	12.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	60.20
GPT4RoI	0.00	0.00	0.00	0.00	10.77	0.00	0.00	0.00	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	40.20	16.07	22.60	0.00	17.83	0.00	0.51	68.40	0.43
LLaVA-NeXT-34B	52.60	22.65	24.35	0.00	14.75	0.12	0.77	72.40	0.67
Pixtral-12B	51.40	32.78	17.4‘	0.00	26.12	0.06	1.16	70.80	0.64
SEED-LLaMA-13B	32.20	4.51	2.17	0.00	1.78	0.00	0.00	30.40	0.00
BLIP2	31.60	40.94	24.82	0.00	29.08	0.00	0.88	67.40	0.57
MiniMonkey	37.00	0.25	0.88	0.00	0.21	0.00	0.00	66.40	0.00
DeepSeek-VL-7B	0.20	0.18	0.00	0.00	1.80	0.00	0.00	0.00	0.20
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	36.30	25.60	27.64	12.07	40.89	13.97	7.55	35.65	25.95
ShareGPT4V-7B	29.86	15.82	19.33	11.63	33.60	13.18	5.58	32.50	16.51
ShareGPT4V-13B	32.91	23.36	19.63	11.00	40.88	13.60	6.07	35.32	20.79

Model	#L-17 (Affect Computing)								
	#10↑	#11↑	#12↑	#13↑	#14↑	#15↑	#16↑	#17↑	#18↑
BLIP-3 (XGen-MM)	39.56	20.57	19.83	14.77	38.76	17.31	5.44	37.51	19.70
AnyGPT	14.45	8.84	8.39	4.93	15.69	6.82	3.98	15.07	10.04
MiniCPM3-4B	39.65	26.91	28.14	22.07	37.25	15.44	10.46	31.81	28.96
LaVIT-V2 (7B)	27.53	20.62	20.58	13.19	31.68	13.61	8.48	20.76	28.42
GLM-VL-Chat	39.69	26.90	26.00	20.48	37.62	12.36	9.68	28.40	28.79
Gemini-1.5-Pro	67.20	45.61	47.57	2.02	51.96	4.78	6.17	75.20	0.60
Gemini-1.5-Flash	58.40	35.49	38.14	1.24	43.26	2.39	4.08	69.60	0.40
OMG-LLaVA-InternLM20B	9.60	1.42	0.88	0.00	1.10	0.00	0.00	20.20	0.00
Idefics3-8B-Llama3	42.20	28.71	28.65	2.28	33.57	2.54	2.25	63.40	0.00
Yi-Vision-v2	38.60	32.69	32.36	0.00	37.49	0.24	3.18	71.43	0.47
Qwen2-VL-72B	69.34	55.78	61.37	0.00	78.03	2.49	10.18	75.81	2.59
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.67	0.00
InternVL2-26B	48.80	34.10	35.40	0.00	38.90	0.00	0.00	63.40	0.00
Qwen2-VL-72B	73.20	55.30	60.60	0.10	74.10	0.30	0.90	76.60	0.30
DeepSeek-VL-2-small	41.60	11.50	13.30	0.00	27.80	0.10	1.10	67.00	0.10
DeepSeek-VL-2	34.20	26.10	32.00	0.00	26.60	0.00	0.00	69.60	0.40
LLaVA-One-Vision-7B	57.00	14.50	10.70	0.00	16.60	0.30	0.90	65.80	0.20
LLaVA-One-Vision-72B	0.20	53.90	50.40	0.00	58.60	2.50	11.30	60.20	1.50
Sa2VA-8B	60.60	27.20	33.10	0.00	51.20	0.70	1.00	69.80	0.70
Sa2VA-26B	60.80	37.40	36.00	0.00	36.50	0.20	5.40	73.60	1.30
CoLVA-2B	23.80	0.50	2.90	0.00	21.20	0.00	0.00	62.00	0.00
CoLVA-4B	22.00	28.90	29.50	0.00	34.50	1.70	1.30	75.60	0.30
Long-LLaVA	42.60	1.50	0.40	0.00	1.00	0.00	0.00	54.00	0.00
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	0.00	8.12	7.24	13.57	0.00	3.01	4.79	35.62	0.00
PandaGPT (13B)	0.00	0.04	0.00	0.09	0.00	0.01	0.00	0.00	0.00
AnyGPT	8.84	8.39	4.93	15.69	6.82	3.98	15.07	10.04	7.82
GAMA	0.00	0.08	0.00	0.09	0.00	0.02	0.00	34.70	0.00
Pengi	0.00	0.04	0.00	0.05	0.00	0.01	0.00	23.50	0.00
SALMONN-7B	1.20	0.00	0.00	0.00	1.17	0.00	0.00	5.60	3.40
SALMONN-13B	1.40	0.04	0.00	0.09	1.19	0.01	0.00	5.90	0.00
WavLLM	0.00	0.07	0.00	0.06	0.00	0.02	0.00	33.78	0.00
ImageBind-LLM	0.00	0.04	0.00	0.09	0.00	0.01	0.00	0.00	0.00
Unified-io-2-XXL	0.00	0.00	0.00	0.00	4.67	0.00	0.00	0.00	0.00
ModaVerse-7b-v0	0.00	0.04	0.00	0.09	0.00	0.01	0.00	0.00	0.00
AudioGPT-GPT4	66.40	41.05	51.08	0.00	67.96	0.80	12.01	72.00	2.31
SpeechGPT-7B-com	0.00	0.08	0.00	0.09	0.00	0.02	0.00	34.70	0.00
LLaMA-Omni	16.40	18.50	5.40	1.56	0.00	29.25	1.10	0.33	2.02
3D-LLM	5.80	0.00	0.00	0.00	0.00	0.00	0.00	65.60	0.00
PointLLM-7B	0.00	0.09	0.00	0.10	0.00	0.03	0.00	0.00	0.00
PointLLM-13B	0.00	0.09	0.00	0.10	0.00	0.03	0.00	0.00	0.00
3D-VisTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-T5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-LLaMA	0.10	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
AvatarGPT	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaMA-Mesh	6.12	0.00	1.62	0.89	0.07	2.21	0.05	0.05	0.00

Table 85: Results on NLP Group, #L-18.

Model	#L-18 (NER)												
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑	#10↑	#11↑	#12↑	#13↑
SoTA Specialist	83.51	74.45	67.00	76.02	92.00	88.08	94.21	97.65	96.74	81.97	78.21	89.18	82.32
Meta-Llama-3.1-8B-Instruct	6.81	17.76	15.53	5.86	14.09	18.77	25.43	30.05	51.57	72.00	12.70	31.22	16.62
Qwen2.5-7B-Instruct	40.07	26.85	25.67	27.54	26.97	36.38	56.83	79.95	45.58	17.98	40.24	23.05	18.58
Gemma-2-9b-it	43.03	32.32	26.62	38.48	33.81	39.95	73.65	82.90	78.18	18.78	43.01	28.66	34.53
ChatGLM-6b	0.05	0.13	0.03	0.00	0.01	0.00	0.18	0.00	0.00	0.00	0.39	0.57	0.15
Vicuna-7b-v1.5	0.04	0.00	0.02	0.00	0.00	0.56	0.00	0.17	0.00	0.00	0.00	0.03	0.13
InternLM-Chat-7b	1.65	0.12	0.47	1.20	1.37	0.72	0.19	13.63	0.00	0.52	12.40	1.31	0.78
GPT-J-6B	0.04	0.00	0.01	0.00	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.03	0.13
Falcon3-7B-Instruct	30.52	15.97	14.21	22.00	7.11	13.81	30.45	78.19	48.84	11.30	41.06	22.58	23.29
Baichuan2-7B-Base	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ministral-8B-Instruct-2410	6.46	6.20	0.52	4.57	5.11	1.75	2.11	29.85	3.11	3.36	8.65	4.90	5.08
Yi-Ligntning	19.60	38.11	24.46	48.11	25.09	25.29	25.00	31.84	44.80	37.15	1.81	42.17	4.14
GPT-3.5-turbo	40.48	41.50	9.12	21.80	26.01	28.93	0.91	59.32	2.34	14.68	39.67	24.06	27.84
GPT-4v	47.34	57.67	29.84	50.56	70.18	49.48	61.78	96.06	87.41	23.94	50.49	41.81	37.24
GPT-4o	60.84	48.43	36.26	50.89	74.90	50.71	82.38	94.89	85.62	29.12	49.34	46.08	37.00
GPT4o-mini	53.21	39.40	33.52	42.35	40.03	48.17	78.55	95.78	93.14	27.61	42.31	37.98	35.95
GPT-4o-4096	60.07	51.72	34.95	50.73	72.61	47.95	81.52	96.30	77.38	28.60	49.36	45.90	36.41
ChatGPT-4o-latest	55.99	46.35	35.81	50.06	72.10	47.39	76.72	95.28	56.99	28.39	49.87	45.55	33.16
Claude-3.5-Sonnet	28.75	21.09	20.31	29.93	24.02	25.56	30.27	41.19	36.19	19.92	23.67	34.53	26.76
Claude-3.5-Opus	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74
Emu2-32B	16.88	14.72	14.19	21.02	17.85	16.18	24.75	32.26	28.44	13.39	18.51	23.37	16.32
DetGPT	10.75	15.19	9.71	12.47	15.07	10.30	15.66	29.04	23.02	6.23	16.82	18.50	13.25
InternVL2.5-8B	7.37	30.33	17.51	33.97	7.54	15.29	9.72	21.90	33.78	13.19	13.52	73.71	9.71
InternVL2.5-4B	8.71	28.29	18.36	26.35	17.75	13.19	6.16	23.17	17.07	14.97	58.62	78.65	55.51
NExT-GPT-V1.5	9.04	0.00	0.02	0.00	0.00	0.07	5.85	0.17	8.53	12.46	8.45	3.03	10.13
InternVL2.5-2B	0.65	1.36	2.09	9.17	1.21	1.13	2.03	2.91	1.27	0.52	0.47	9.23	1.74
Monkey-10B-chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00
DeepSeek-VL-7B	0.25	0.81	1.86	5.42	1.05	0.65	1.34	2.07	4.07	1.51	2.32	30.05	2.19
Qwen2-VL-7B	0.00	0.58	0.00	0.00	0.00	0.42	0.00	1.08	0.00	0.00	0.35	0.00	0.65
Qwen-VL-Chat	1.00	0.27	0.11	0.22	0.47	0.70	1.78	2.19	0.09	0.09	0.83	0.08	0.81
Qwen-Audio-Chat	1.58	4.47	1.03	2.65	0.45	2.56	10.32	7.60	16.78	0.13	12.45	2.89	1.45
Qwen2-Audio-Instruct	1.59	4.41	1.22	2.92	0.62	3.35	11.62	8.10	18.72	0.27	13.19	2.97	2.21
MoE-LLAVA-Phi2-2.7B-4e-384	1.13	1.42	1.70	1.96	0.80	2.00	12.35	25.27	10.40	0.23	2.87	0.36	1.83
mPLUG-Owl2-LLaMA2-7b	0.08	0.10	0.04	1.72	0.20	0.56	2.38	11.95	0.87	0.32	2.54	0.17	0.27
Phi-3.5-Vision-Instruct	5.41	3.86	0.41	3.73	0.90	3.31	23.37	61.43	29.48	1.73	21.49	6.07	5.61
Cambrian-1-8B	0.30	4.95	0.55	2.45	1.30	2.56	0.00	2.44	0.00	0.00	3.53	1.27	0.32
MiniGPT4-LLaMA2	8.34	1.54	2.07	6.74	12.56	3.49	7.42	15.11	7.58	3.27	8.67	3.44	2.45
InternVL-Chat-V1-5	6.38	29.98	15.53	31.03	8.93	13.62	4.20	22.11	14.35	7.80	12.32	70.80	14.85
Mini-InternVL-Chat-4B-V1-5	3.11	16.73	5.84	20.87	4.56	10.52	5.20	15.31	5.30	9.81	32.76	61.38	26.63
InternLM-XComposer2-VL-1.8B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	5.01	5.43	2.28	3.34	0.68	2.59	22.10	52.72	31.21	0.22	3.88	1.34	3.08
LLaVA-NeXT-34B	7.45	7.92	5.13	4.98	1.75	3.68	26.94	61.26	32.22	0.74	8.85	2.34	2.42
Pixtral-12B	8.23	6.65	4.74	13.72	0.69	4.74	25.01	63.06	34.76	0.82	14.16	5.72	4.76
SEED-LLaMA-13B	1.74	0.65	0.66	2.85	0.10	0.57	4.82	16.58	4.76	0.03	0.42	0.18	0.25
BLIP2	1.42	3.49	0.45	16.51	0.00	2.17	4.56	35.70	19.64	0.00	7.83	0.18	0.31
MiniMonkey	0.32	0.79	0.21	2.78	0.00	0.38	5.74	27.81	25.35	0.21	3.46	0.22	0.32
DeepSeek-VL-7B	0.00	0.05	0.09	0.12	0.05	0.07	0.00	0.00	0.00	0.27	1.14	0.61	0.44
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	17.28	21.47	9.66	19.91	20.38	14.37	25.75	38.26	26.80	13.30	22.25	23.08	20.29
ShareGPT4V-7B	10.53	14.50	5.39	14.02	14.62	12.19	14.38	31.20	21.91	4.94	15.69	18.11	8.88
ShareGPT4V-13B	17.46	17.27	9.05	17.92	15.32	14.32	17.51	31.09	23.35	11.16	20.00	20.46	16.81

Model	#L-18 (NER)												
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑	#10↑	#11↑	#12↑	#13↑
BLIP-3 (XGen-MM)	18.38	15.56	13.15	16.92	16.46	13.65	18.39	33.99	29.00	9.63	18.34	22.41	12.03
AnyGPT	7.82	8.57	5.26	0.00	6.02	5.71	10.10	10.93	10.35	6.15	9.54	9.06	7.64
MiniCPM3-4B	17.83	15.66	14.07	19.69	16.49	12.26	26.47	29.23	26.81	10.83	19.29	31.37	17.12
LaVIT-V2 (7B)	11.16	14.05	9.28	11.98	9.97	11.13	11.15	23.96	15.05	9.25	18.54	26.57	17.22
GLM-VL-Chat	16.49	17.01	11.54	17.76	13.38	10.06	23.00	28.73	24.81	8.91	20.49	31.68	16.10
Gemini-1.5-Pro	48.53	50.66	29.19	43.31	49.74	46.51	67.45	90.12	82.53	19.62	47.10	42.75	36.80
Gemini-1.5-Flash	41.72	42.52	27.37	37.98	40.29	42.16	60.03	82.95	79.40	16.23	43.91	38.49	32.13
OMG-LLaVA-InternLM20B	4.55	7.28	2.20	4.43	2.26	3.80	6.82	11.64	13.71	0.00	3.47	2.05	2.33
Idefics3-8B-Llama3	27.46	32.75	21.06	28.65	26.51	25.56	45.38	73.20	66.23	14.18	19.59	25.97	26.54
Yi-Vision-v2	25.61	19.34	5.84	22.92	21.08	7.48	38.30	76.28	58.70	13.61	35.33	17.81	23.02
Qwen2-VL-72B	50.59	56.57	23.52	40.31	48.31	40.65	85.92	93.81	87.47	18.33	52.48	43.63	35.58
Otter	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2-26B	29.30	19.90	0.60	25.20	20.70	15.00	0.90	73.30	21.70	0.70	34.50	13.10	17.30
Qwen2-VL-72B	47.00	52.20	26.20	43.10	54.20	45.10	84.90	94.20	84.90	19.00	49.20	40.00	33.20
DeepSeek-VL-2-small	5.50	12.40	0.10	4.40	3.50	1.20	15.70	55.70	12.90	3.30	16.10	2.60	3.40
DeepSeek-VL-2	20.90	14.60	1.70	15.50	10.30	3.80	13.70	80.60	15.30	3.70	21.80	3.70	1.70
LLaVA-One-Vision-7B	3.10	3.80	2.70	11.90	5.40	7.30	0.80	2.70	0.10	1.00	11.60	1.40	2.70
LLaVA-One-Vision-72B	44.40	32.70	11.90	35.00	62.00	11.10	71.30	91.00	81.10	19.20	32.60	37.10	28.80
Sa2VA-8B	29.50	13.50	8.00	22.70	38.60	16.90	24.50	73.50	0.00	7.70	45.70	4.80	6.90
Sa2VA-26B	32.20	21.40	11.30	25.20	6.70	22.90	60.00	80.50	4.40	10.80	42.20	15.30	18.30
CoLVA-2B	3.00	1.20	0.40	0.70	2.80	0.80	5.20	17.60	0.00	0.20	2.20	0.50	0.10
CoLVA-4B	29.50	6.80	2.70	14.40	18.50	12.70	65.60	77.50	58.80	4.40	23.90	11.30	4.60
Long-LLaVA	0.10	0.20	0.10	1.20	0.10	0.60	1.10	26.20	0.10	0.00	0.40	0.00	0.20
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	10.03	0.00	0.02	0.00	0.00	0.07	9.08	0.17	10.30	10.35	16.74	5.03	8.19
PandaGPT (13B)	0.04	0.00	0.02	0.00	0.00	0.56	0.00	0.17	0.00	0.00	0.00	0.03	0.13
AnyGPT	8.57	5.26	0.00	6.02	5.71	10.10	10.93	10.35	6.15	9.54	9.06	7.64	10.00
GAMA	0.04	0.00	0.02	0.00	0.00	0.58	5.93	0.17	8.05	0.00	0.00	0.03	0.13
Pengi	0.04	0.00	0.02	0.00	0.00	0.13	0.00	0.15	0.00	0.00	0.00	0.01	0.08
SALMONN-7B	0.00	0.14	0.00	0.11	0.00	0.21	1.67	4.74	0.00	0.00	0.86	0.00	0.31
SALMONN-13B	0.04	0.18	0.02	0.00	0.00	0.56	1.89	5.17	0.00	0.00	0.00	0.03	0.13
WavLLM	0.03	0.00	0.02	0.00	0.00	0.07	5.69	0.23	8.23	0.00	0.00	0.03	0.12
ImageBind-LLM	0.04	0.00	0.02	0.00	0.00	0.56	0.00	0.17	0.00	0.00	0.00	0.03	0.13
Unified-io-2-XXL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ModaVerse-7b-v0	0.04	0.00	0.02	0.00	0.00	0.56	0.00	0.17	0.00	0.00	0.00	0.03	0.13
AudioGPT-GPT4	46.90	56.18	29.14	49.32	78.34	45.90	61.78	95.67	85.41	20.56	23.45	40.81	35.23
SpeechGPT-7B-com	0.04	0.00	0.02	0.00	0.00	0.07	5.85	0.17	8.53	0.00	0.00	0.03	0.13
LLaMA-Omni	6.23	14.67	15.40	5.86	12.40	14.30	23.50	28.95	13.45	0.00	0.00	5.22	4.34
3D-LLM	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.16	0.00	0.00	0.24
PointLLM-7B	0.04	0.00	0.02	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.00	0.03	0.16
PointLLM-13B	0.04	0.00	0.02	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.00	0.03	0.16
3D-VisTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-T5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-LLaMA	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.04	0.00	0.00
AvatarGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaMA-Mesh	0.00	1.05	0.19	0.06	1.24	0.57	0.94	0.62	2.49	0.17	0.06	0.68	0.44

Table 87: Results on NLP Group, #L-19.

Model	#L-19 (Cog QA)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
SoTA Specialist	91.87	95.27	83.64	75.57	91.61	65.23	89.20	72.40	88.20
Meta-Llama-3.1-8B-Instruct	43.70	38.80	70.40	0.00	21.74	7.65	42.00	46.80	5.20
Qwen2.5-7B-Instruct	59.80	72.20	0.00	29.29	83.00	10.80	28.20	48.40	8.60
Gemma-2-9b-it	64.40	76.20	0.00	39.48	68.60	6.95	43.20	48.20	15.60
ChatGLM-6b	6.40	46.00	0.00	3.52	0.20	0.00	1.20	16.80	1.60
Vicuna-7b-v1.5	0.00	0.00	0.00	4.26	0.00	0.00	0.00	0.00	0.00
InternLM-Chat-7b	8.20	27.00	0.00	4.94	21.60	1.19	23.00	20.80	5.00
GPT-J-6B	0.00	0.00	0.00	3.86	0.00	0.00	0.00	0.00	0.00
Falcon3-7B-Instruct	51.00	70.20	0.00	40.25	74.80	6.97	6.80	47.60	10.40
Baichuan2-7B-Base	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00
Ministral-8B-Instruct-2410	47.60	62.60	0.00	14.00	63.40	5.85	40.00	42.80	6.20
Yi-Ligntning	76.80	77.60	0.00	45.33	72.40	6.19	44.80	49.40	13.00
GPT-3.5-turbo	52.00	81.80	0.00	29.30	80.00	8.18	43.00	48.40	11.20
GPT-4v	70.00	86.20	0.00	67.28	77.40	9.78	45.80	47.80	4.80
GPT-4o	83.20	85.90	0.00	63.34	76.80	8.69	45.80	48.80	10.20
GPT4o-mini	69.80	83.00	0.00	35.21	75.79	9.44	43.20	48.80	15.60
GPT-4o-4096	79.60	86.20	0.00	60.19	75.80	9.87	46.20	48.20	5.00
ChatGPT-4o-latest	80.20	86.80	0.00	59.31	73.80	8.97	44.60	48.40	8.00
Claude-3.5-Sonnet	41.60	57.67	6.36	24.33	66.38	11.04	34.29	39.44	11.05
Claude-3.5-Opus	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74
Emu2-32B	30.80	44.61	0.00	20.31	56.97	5.94	22.64	34.82	0.00
DetGPT	23.02	41.98	0.00	16.31	50.67	2.95	15.86	32.02	0.00
InternVL2.5-8B	49.40	76.20	0.00	27.01	77.20	6.93	38.00	48.00	5.60
InternVL2.5-4B	40.00	67.00	0.00	19.77	72.00	4.68	30.20	46.80	7.60
NExT-GPT-V1.5	19.05	4.32	8.59	5.43	2.30	9.44	6.32	0.00	0.00
InternVL2.5-2B	19.60	36.00	0.00	11.90	71.00	1.31	3.00	43.40	4.40
Monkey-10B-chat	3.00	0.80	0.00	3.49	11.20	0.00	1.20	7.40	0.00
DeepSeek-VL-7B	39.00	64.60	0.00	5.27	73.20	4.02	23.00	46.80	0.21
Qwen2-VL-7B	0.00	1.00	0.00	4.20	0.00	1.18	11.83	25.60	0.00
Qwen-VL-Chat	1.00	0.00	0.00	4.60	0.00	0.12	13.08	26.91	0.08
Qwen-Audio-Chat	13.50	49.78	0.00	5.34	0.80	5.34	6.30	27.50	3.20
Qwen2-Audio-Instruct	15.20	50.60	0.00	5.52	1.00	5.27	6.20	30.00	3.80
MoE-LLAVA-Phi2-2.7B-4e-384	1.13	53.60	0.00	3.73	64.00	6.09	4.47	16.47	0.36
mPLUG-Owl2-LLaMA2-7b	0.08	20.60	0.00	2.06	31.40	0.16	8.20	21.21	0.17
Phi-3.5-Vision-Instruct	5.41	75.20	0.00	4.17	68.80	6.13	9.31	22.83	6.07
Cambrian-1-8B	0.30	5.80	0.00	2.28	83.40	3.67	0.00	16.70	1.27
MiniGPT4-LLaMA2	8.34	0.00	0.00	3.96	0.00	0.78	4.86	19.31	3.44
InternVL-Chat-V1-5	53.20	66.80	0.00	32.30	69.80	4.64	19.00	42.20	4.00
Mini-InternVL-Chat-4B-V1-5	27.00	50.80	0.00	12.96	66.60	5.46	33.60	48.00	5.20
InternLM-XComposer2-VL-1.8B	5.00	17.60	0.00	0.00	46.80	0.00	85.20	27.60	2.80
GPT4RoI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-13B	17.77	49.94	0.00	17.63	63.20	5.36	18.80	37.40	4.40
LLaVA-NeXT-34B	21.08	63.60	0.00	33.85	67.40	6.47	17.40	47.60	5.00
Pixtral-12B	23.71	61.74	0.00	32.45	72.20	7.12	20.20	48.80	5.40
SEED-LLaMA-13B	1.03	14.22	0.00	7.55	34.60	0.76	7.60	26.20	0.80
BLIP2	23.31	41.16	0.00	8.77	57.60	2.04	19.80	46.20	1.00
MiniMonkey	16.20	36.60	0.00	9.63	62.20	0.18	16.40	48.60	0.60
DeepSeek-VL-7B	0.00	0.20	2.00	2.05	14.20	0.00	0.00	0.00	4.00
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	26.85	50.96	0.00	19.45	53.17	6.88	20.10	36.22	4.88
ShareGPT4V-7B	26.63	43.07	0.00	15.45	49.86	1.22	17.29	31.63	1.56
ShareGPT4V-13B	26.69	42.06	0.00	17.01	53.13	3.64	19.02	33.37	2.05

Model	#L-19 (Cog QA)								
	#1↑	#2↑	#3↑	#4↑	#5↑	#6↑	#7↑	#8↑	#9↑
BLIP-3 (XGen-MM)	31.79	49.11	0.00	17.94	54.09	4.65	20.88	34.96	5.53
AnyGPT	10.00	18.03	3.13	7.06	18.88	3.42	8.66	11.78	0.00
MiniCPM3-4B	27.08	46.04	6.44	17.45	59.27	13.48	24.74	36.41	15.05
LaVIT-V2 (7B)	20.80	39.84	3.73	10.97	41.93	5.39	14.56	26.45	9.32
GLM-VL-Chat	25.03	45.62	7.49	15.81	59.68	13.30	23.83	37.89	15.67
Gemini-1.5-Pro	52.34	85.00	0.00	36.45	75.24	7.30	44.20	46.80	6.00
Gemini-1.5-Flash	48.95	79.53	0.00	32.93	70.45	6.96	40.80	39.40	7.00
OMG-LLaVA-InternLM20B	5.26	10.14	0.00	3.10	28.71	0.00	7.60	9.40	0.00
Idefics3-8B-Llama3	38.09	62.18	0.00	22.93	69.20	5.24	31.00	40.20	4.80
Yi-Vision-v2	53.80	50.03	0.00	12.43	72.20	6.61	42.80	27.25	8.40
Qwen2-VL-72B	63.33	71.85	0.00	61.37	77.33	18.97	16.00	58.67	11.33
Otter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Show-o	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NExT-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2-26B	60.40	78.00	0.00	31.70	70.80	0.70	41.40	45.20	0.60
Qwen2-VL-72B	66.60	83.40	0.00	38.60	75.40	0.90	43.60	48.20	0.80
DeepSeek-VL-2-small	23.80	31.00	0.00	7.80	67.20	5.40	10.40	45.60	11.40
DeepSeek-VL-2	33.00	72.80	0.00	1.70	72.00	6.40	44.40	48.60	1.40
LLaVA-One-Vision-7B	48.80	68.00	0.00	24.10	74.60	4.20	8.60	45.60	7.80
LLaVA-One-Vision-72B	60.00	76.00	0.00	62.70	0.00	7.60	44.00	0.20	7.20
Sa2VA-8B	55.60	78.80	0.00	13.30	80.80	9.40	42.80	47.40	1.60
Sa2VA-26B	72.40	76.40	0.00	46.80	75.00	6.40	46.00	45.80	9.80
CoLVA-2B	14.00	7.40	0.00	3.30	42.00	1.20	2.40	22.80	3.00
CoLVA-4B	38.20	66.20	0.00	27.10	75.60	8.80	17.00	47.20	10.60
Long-LLaVA	9.40	51.60	0.00	18.20	80.00	5.70	2.80	22.00	4.00
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	32.16	14.35	8.59	5.43	3.40	9.54	3.63	0.00	0.00
PandaGPT (13B)	0.00	0.00	0.00	2.76	0.00	0.00	0.00	0.00	0.00
AnyGPT	18.03	3.13	7.06	18.88	3.42	8.66	11.78	0.00	4.43
GAMA	0.00	0.00	0.00	4.23	0.00	0.00	0.00	0.00	0.00
Pengi	0.00	0.00	0.00	2.16	0.00	0.00	0.00	0.00	0.00
SALMONN-7B	0.40	0.20	0.00	0.25	0.80	0.00	4.80	3.40	8.20
SALMONN-13B	0.80	0.40	0.00	0.76	1.10	0.00	5.00	3.60	8.50
WavLLM	0.00	0.00	0.00	4.57	0.00	0.00	0.00	0.00	0.00
ImageBind-LLM	0.00	0.00	0.00	2.76	0.00	0.00	0.00	0.00	0.00
Unified-io-2-XXL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ModaVerse-7b-v0	0.00	0.00	0.00	2.76	0.00	0.00	0.00	0.00	0.00
AudioGPT-GPT4	71.34	83.45	0.00	56.80	73.20	9.45	42.10	45.35	4.30
SpeechGPT-7B-com	0.00	0.00	0.00	4.65	0.00	0.00	0.00	0.00	0.00
LLaMA-Omni	0.00	0.00	0.00	0.00	20.56	7.85	1.80	4.60	4.60
3D-LLM	3.80	0.00	0.00	7.90	0.40	0.00	0.00	0.00	0.00
PointLLM-7B	0.00	0.00	0.00	4.26	0.00	0.00	0.00	0.00	0.00
PointLLM-13B	0.00	0.00	0.00	4.26	0.00	0.00	0.00	0.00	0.00
3D-VisTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-T5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-LLaMA	0.18	0.00	0.00	0.00	1.19	0.00	0.00	0.00	0.00
AvatarGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaMA-Mesh	0.27	0.00	0.00	0.00	4.68	0.00	0.10	0.00	0.00

Table 89: Results on NLP Group, from #L-20 to #L-22.

Model	#L-20 (Event Ext)		#L-21 (Sem Par)			#L-22 (Ling Par)		
	#1↑	#2↑	#1↑	#2↑	#3↑	#1↑	#2↑	#3↑
SoTA Specialist	62.89	54.32	69.00	89.40	74.80	93.30	91.40	92.43
Meta-Llama-3.1-8B-Instruct	1.36	14.80	34.00	38.00	25.20	57.80	36.80	68.46
Qwen2.5-7B-Instruct	3.67	27.80	51.20	57.60	29.40	78.40	49.60	44.49
Gemma-2-9b-it	18.60	20.20	48.60	47.80	30.20	84.40	55.40	72.27
ChatGLM-6b	0.00	5.40	28.20	20.40	22.80	64.60	50.60	20.91
Vicuna-7b-v1.5	0.00	0.00	26.20	20.40	22.80	24.20	21.20	0.79
InternLM-Chat-7b	0.02	6.60	35.80	24.00	21.40	54.40	31.00	21.86
GPT-J-6B	0.00	0.00	26.20	20.40	22.80	24.20	21.20	23.80
Falcon3-7B-Instruct	3.77	18.60	42.60	45.20	31.60	79.80	48.00	48.46
Baichuan2-7B-Base	0.00	0.00	6.15	0.00	0.00	0.00	0.00	0.00
Ministral-8B-Instruct-2410	0.33	18.40	28.00	24.40	24.20	53.80	34.40	33.12
Yi-Ligntning	21.67	29.00	38.40	20.80	28.60	83.60	67.00	30.88
GPT-3.5-turbo	15.30	19.20	0.00	0.00	0.00	0.00	0.00	70.06
GPT-4v	35.72	32.00	32.00	42.99	4.40	0.20	3.80	68.72
GPT-4o	42.03	37.00	37.00	45.60	13.60	1.40	5.40	78.69
GPT4o-mini	25.02	25.00	25.00	53.60	8.80	0.60	6.80	70.90
GPT-4o-4096	39.20	38.80	60.80	87.20	36.40	86.60	1.00	56.27
ChatGPT-4o-latest	38.29	32.40	59.60	79.80	25.60	61.40	0.00	41.54
Claude-3.5-Sonnet	14.30	25.37	36.00	43.00	28.60	33.40	30.20	39.97
Claude-3.5-Opus	30.74	30.74	30.74	30.74	30.74	30.74	30.74	30.74
Emu2-32B	10.58	16.97	25.60	28.60	25.20	30.40	27.40	37.35
DetGPT	2.19	10.62	24.40	29.40	18.20	23.20	21.20	28.58
InternVL2.5-8B	1.20	18.40	40.59	30.40	25.40	66.20	41.99	52.82
InternVL2.5-4B	2.87	16.20	37.80	24.80	25.60	39.40	38.80	34.23
NExT-GPT-V1.5	4.63	9.80	28.95	21.45	22.10	32.70	19.50	4.37
InternVL2.5-2B	0.00	7.80	29.60	32.80	24.60	53.20	26.60	25.63
Monkey-10B-chat	0.00	1.40	5.20	20.00	22.40	19.20	10.00	0.00
DeepSeek-VL-7B	0.21	16.00	34.80	36.60	24.00	59.60	39.00	24.30
Qwen2-VL-7B	0.00	12.00	19.06	20.17	23.38	21.00	29.00	15.38
Qwen-VL-Chat	0.01	0.64	20.80	20.31	19.63	18.59	21.20	9.12
Qwen-Audio-Chat	0.00	4.60	30.40	23.10	22.50	49.60	38.70	20.30
Qwen2-Audio-Instruct	0.00	5.00	31.80	24.00	23.60	50.80	39.60	21.11
MoE-LLAVA-Phi2-2.7B-4e-384	0.00	9.78	16.40	19.80	21.00	20.30	40.80	23.77
mPLUG-Owl2-LLaMA2-7b	0.06	6.52	16.79	18.50	21.01	19.43	25.40	23.83
Phi-3.5-Vision-Instruct	0.03	21.83	15.82	20.06	25.37	21.30	33.00	19.64
Cambrian-1-8B	0.00	8.56	12.86	21.60	17.88	20.74	2.40	22.24
MiniGPT4-LLaMA2	0.00	0.00	13.40	18.94	20.32	17.55	24.00	22.15
InternVL-Chat-V1-5	0.00	11.80	32.80	27.60	26.00	61.60	40.00	43.23
Mini-InternVL-Chat-4B-V1-5	0.00	15.80	33.40	27.60	23.40	55.20	30.00	36.06
InternLM-XComposer2-VL-1.8B	0.00	0.00	31.60	34.00	31.60	61.60	38.20	18.72
GPT4RoI	0.00	0.00	1.40	9.00	22.40	6.00	10.20	0.00
GLaMM	0.00	0.00	0.00	0.00	0.00	1.40	0.00	0.00
LLaVA-NeXT-13B	0.00	16.20	28.80	25.60	22.20	21.20	29.20	23.45
LLaVA-NeXT-34B	0.00	17.20	30.60	28.60	20.40	32.20	33.40	22.17
Pixtral-12B	0.46	24.80	33.80	32.80	26.20	24.60	31.20	28.89
SEED-LLaMA-13B	0.00	4.20	11.60	17.80	6.80	15.80	14.20	7.23
BLIP2	0.33	5.80	33.60	25.80	21.40	49.80	33.60	29.06
MiniMonkey	0.00	3.60	30.20	26.00	21.80	51.60	31.80	20.80
DeepSeek-VL-7B	0.00	0.00	22.40	20.80	22.60	24.20	26.40	24.79
LISA-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CogVLM-Chat	13.80	19.88	26.40	34.00	20.00	27.00	24.20	33.50
ShareGPT4V-7B	5.70	13.20	20.40	28.00	20.80	20.60	20.40	27.92
ShareGPT4V-13B	7.48	14.39	25.20	32.80	19.40	25.20	21.80	32.65

Model	#L-20 (Event Ext)		#L-21 (Sem Par)			#L-22 (Ling Par)		
	#1↑	#2↑	#1↑	#2↑	#3↑	#1↑	#2↑	#3↑
BLIP-3 (XGen-MM)	6.27	14.19	26.80	31.60	22.20	25.00	23.40	33.68
AnyGPT	4.43	5.69	9.60	11.20	10.20	11.00	9.80	11.72
MiniCPM3-4B	13.48	21.15	29.80	35.80	28.60	31.80	35.60	41.51
LaVIT-V2 (7B)	7.11	14.27	24.80	24.20	24.60	27.40	26.40	28.87
GLM-VL-Chat	10.77	22.53	29.00	32.60	24.00	29.60	33.20	40.37
Gemini-1.5-Pro	22.16	28.62	48.60	56.40	29.40	83.80	73.20	71.42
Gemini-1.5-Flash	13.45	22.31	43.20	49.20	27.20	82.60	66.20	62.03
OMG-LLaVA-InternLM20B	0.00	2.40	13.60	10.40	6.80	9.00	5.60	11.67
Idefics3-8B-Llama3	13.92	18.00	35.80	43.80	28.20	69.40	68.60	58.39
Yi-Vision-v2	7.89	15.60	12.40	2.40	25.87	6.20	0.60	41.90
Qwen2-VL-72B	27.94	23.33	55.33	62.00	30.32	87.33	0.00	70.02
Otter	0.00	0.00	26.67	12.67	17.62	18.00	19.33	17.69
Show-o	0.00	0.00	0.00	0.00	0.00	5.30	1.20	7.84
NExT-Chat	0.00	0.00	2.00	4.66	7.91	1.33	24.66	3.43
InternVL2-26B	0.00	12.40	36.80	38.20	25.20	73.80	37.40	41.90
Qwen2-VL-72B	24.80	23.60	54.60	73.40	38.10	89.80	71.60	73.20
DeepSeek-VL-2-small	0.30	15.00	30.60	30.60	24.40	53.20	46.20	28.00
DeepSeek-VL-2	0.20	10.60	35.80	28.00	24.80	62.40	41.00	41.60
LLaVA-One-Vision-7B	0.10	14.60	42.00	31.00	27.20	71.40	46.00	34.30
LLaVA-One-Vision-72B	8.30	30.20	46.00	74.20	36.40	91.40	67.40	57.00
Sa2VA-8B	1.90	19.20	45.40	20.60	27.80	75.00	38.40	49.70
Sa2VA-26B	2.80	18.60	43.00	20.40	28.40	81.80	49.20	43.60
CoLVA-2B	0.00	1.20	25.60	20.20	22.40	44.60	37.40	12.60
CoLVA-4B	2.60	13.80	38.60	30.60	26.40	76.80	37.00	31.90
Long-LLaVA	0.00	14.80	28.00	35.20	25.20	58.00	43.20	25.00
LM4LV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vitron-V1	3.62	9.60	24.36	19.37	22.30	31.60	20.70	5.23
PandaGPT (13B)	0.00	0.00	13.60	25.80	12.50	16.30	25.70	1.45
AnyGPT	5.69	9.60	11.20	10.20	11.00	9.80	11.72	2.57
GAMA	0.00	0.00	15.67	19.86	24.10	18.45	21.32	4.35
Pengi	0.00	0.00	13.60	25.80	12.50	10.75	16.98	0.65
SALMONN-7B	0.00	0.00	15.20	4.80	20.00	16.80	7.60	11.98
SALMONN-13B	0.00	0.00	15.60	5.80	12.50	17.30	7.90	12.66
WavLLM	0.00	0.00	15.69	20.81	24.32	19.54	20.23	3.45
ImageBind-LLM	0.00	0.00	13.60	25.80	12.50	16.30	25.70	0.66
Unified-io-2-XXL	0.00	5.60	1.50	2.50	15.40	3.10	8.40	0.00
ModaVerse-7b-v0	0.00	0.00	13.60	25.80	12.50	16.30	25.70	0.66
AudioGPT-GPT4	32.37	23.00	31.80	40.22	4.40	0.20	3.40	73.34
SpeechGPT-7B-com	0.00	0.00	27.40	20.80	23.10	23.80	20.10	0.65
LLaMA-Omni	1.23	15.10	13.00	17.80	15.10	26.40	13.40	17.34
3D-LLM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35
PointLLM-7B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
PointLLM-13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3D-VisTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-T5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MotionGPT-LLaMA	0.00	0.00	12.28	19.80	22.80	22.80	22.20	0.00
AvatarGPT	0.00	0.00	0.00	0.40	0.00	0.00	0.80	0.00
LLaMA-Mesh	0.02	0.00	26.20	20.40	22.80	24.20	21.10	19.42

D Discussions and Future Investigation

We propose a leveled evaluation of MLLMs, with a hierarchical framework called General-Level, and a large-scale benchmark dataset General-Bench. Yet we believe several aspects of this work can be further improved. In addition, as the next step to achieve more capable multimodal generalists toward AGI, we believe some points are worth further investigation.

Further refinement of the General-Level framework. Although this framework is a concrete starting point for building true multimodal generalists, there are still areas for improvement, especially in terms of the algorithms. For example, the coordination average used to compute Level-3 in the General-Level framework assumes a balance between the number of comprehension and generation tasks, which is an unrealistic assumption. Additionally, we have relaxed the measurement of synergy by assuming that a model's synergy capability is reflected in its ability to surpass SoTA specialists in task performance, avoiding a direct measurement of the synergy effect. Future work will consider optimizing this aspect to provide a more robust definition.

Expanding the General-Bench dataset to include more comprehensive tasks and modalities. To ensure the evaluation of multimodal generalists is complete and unbiased, the General-Bench dataset should be further expanded. Currently, the dataset is somewhat imbalanced across different modalities and tasks; for example, there is more data for image-related tasks than for audio and 3D modalities, and there are more comprehension tasks than generation tasks. Moreover, multimodality should be more broadly defined to include not just visible information such as language, vision, and sound, but also other signals and types of information. For instance, LLMs' reasoning abilities have been shown to be significantly enhanced through learning code. As multimodal models, it is necessary to support some coding capabilities, which, in turn, could theoretically improve the model's understanding and reasoning in other modalities, such as vision. Finally, our current benchmark primarily considers tasks in which individual modality is operated in isolation. However, in reality, a good multimodal generalist should be capable of modality-switching and modality-interleaved reasoning (in both comprehension and generation) under multi-turn user-machine interactions. In the future, we plan to incorporate tasks and datasets that assess interleaved modality capabilities.

Rethinking Evaluation Paradigm for Model Capabilities. Many current task evaluation methodologies still follow conventional paradigms. In most cases, automatic and scalable evaluation strategies are preferred (we provide a detailed overview in Appendix §B.4). While such approaches may suffice for relatively straightforward tasks—such as multiple-choice or classification—they often fall short when applied to format-free tasks, particularly those involving multimodal generation. For instance, in video or 3D generation, traditional metrics like FID or FVD are increasingly considered inadequate, as they fail to reliably capture the quality and fidelity of the generated content. Consequently, there is a growing reliance on human evaluations. To improve scalability, many recent works have begun employing LLMs to simulate human-level judgment (Zheng et al., 2023). However, this "LLM-as-a-judge" approach introduces challenges in terms of evaluation stability and reproducibility, which remain open research problems. In addition, our current General-Level evaluation framework adopts a single primary metric per task, which may inherently introduce bias. We argue that future evaluations should incorporate multiple complementary metrics to provide a more comprehensive assessment. Lastly, as multimodal generalist models continue to evolve with stronger reasoning capabilities, corresponding benchmarks should also be upgraded to evaluate the interpretability and traceability of their intermediate reasoning processes.

Optimizing model architecture to support more functionalities and modalities with stronger performance. Our above experiments reveal that very few MLLMs currently achieve unified capabilities. Most models support only 1-2 modalities or abilities, which severely limits their qualifications as generalists. Some models, while supporting multiple tasks and modalities, still exhibit limited performance in individual tasks. Future research could focus on optimizing model architectures to support as many functions and modalities as possible while also delivering stronger performance. We note that simply integrating multiple models or modules using an agent-based approach can increase the number of supported functions and modalities but does not necessarily improve performance (i.e., it still cannot surpass individual specialists). A more promising approach may be to leverage the Mixture of Experts (MoE) strategy to construct more unified MLLMs. Recently, more advanced understanding and generation capabilities have also been achieved through some of the latest architectures, such as those that combine autoregressive and diffusion frameworks.

Strengthening synergy capabilities is the key focus. As emphasized in this work, achieving synergy is the fundamental requirement, and it is crucial for ensuring the model has more powerful capabilities (compared to specialists). To accomplish this, several aspects need to be considered. First, at the architectural level, it is necessary to design essential modules or mechanisms that allow the model to flexibly and effectively transfer features learned from different tasks and modalities (Fei

et al., 2024a; Pan et al., 2025). This is key to enabling the "learning by analogy" capability. Second, at the learning level, to achieve stronger capabilities across multiple tasks and modalities, the model must be trained in a way that prevents it from forgetting previously learned knowledge when learning new tasks. Also, recently, by incorporating training techniques such as Reinforcement Learning from Human Feedback (RLHF), models have achieved more powerful reasoning capabilities and improved generalization.

E Author Contribution

All authors contributed to this project in various capacities, including idea conceptualization, data annotation, model implementation, paper writing, and project supervision. To provide a transparent overview, Table 91 summarizes the contributions and responsibilities of all co-authors. Given the extensive scope and workload of this project, we enlisted the help of a large group of contributors. Among them, some individuals made contributions but did not qualify for co-authorship due to partial involvement or insufficient or voluntary contributions. Nevertheless, we acknowledge their efforts and list them in Table 93 to express our gratitude for their support.

Table 91: Detailed author list and contribution statement.

#	Name	Group	Role	Responsible Datasets	Responsible Models
1	Hao Fei	All	1) Project general leader: design and implement the idea, including General-Level evaluation and General-Bench planning. 2) Worker for the audio group. 3) Designed data collection methods and selected models for verification. 4) Responsible for all paper writing, illustrations, and polishing. 5) Managed online deployment of data and automated evaluation systems. 6) Maintained the project website. 7) Provided computing resources.	All audio-generation datasets	All specialists in audio-generation tasks Audio MLLMs: WavLLM, ImageBind-LLM, Unified-io-2-XXL, ModaVerse-7b-v0, AudioGPT-GPT4, SpeechGPT-7B-com, LLaMA-Omni
2	Yuan Zhou	Working for Image group	1) Project co-leader: give the formal text and formula definitions for the 5 levels of the General-Level evaluation framework, along with the corresponding formula derivation. 2) Led the image group, managing tasks and execution. 3) Constructed and polished over 150 datasets; implemented around 30 SoTA specialists and 2 MLLMs. 4) Verified task and data management, and deployed systems. 5) Developed evaluation scripts and automated testing systems.	All image-related datasets	Image-oriented MLLMs: GPT4-o, GPT4-o-mini, GPT4-V
3	Juncheng Li	Working for Image group	Project co-leader for image group: Led the image group for supervised dataset collection and image-based MLLMs evaluation.	All image-related datasets	Specialists and MLLMs supporting image-related skills
4	Xiangtai Li	Working for Video group	Project co-leader for video group: Led the video group for supervised datasets collection and video-based MLLMs evaluation.	All video-related datasets	Specialists and MLLMs supporting video-related skills
5	Qingshan Xu	Working for 3D group	Project co-leader for 3D group: Led the 3D group for supervised datasets collection and 3D-based MLLMs evaluation.	All 3D-related datasets	Specialists and MLLMs supporting 3D-related skills

#	Name	Group	Role	Responsible Datasets	Responsible Models
6	Bobo Li	Working for Language group	Project co-leader for Language group: Led the Language group for supervised datasets collection and Language-based MLLMs evaluation.	Language-related datasets: L-2, 14, 15, 16, 17, 18, 19, 20, 21, 22	Specialists supporting L-2, 14, 15, 16, 17, 18, 19, 20, 21, 22 skills, and MLLMs including Qwen2.5-7B-Instruct, Baichuan2-7B-Base, Vicuna-7b-V1.5, Falcon3-7B-Instruct, Ministral-8B-Instruct-2410
7	Shengqiong Wu	Working for Audio group	Project co-leader for the audio group: Led the audio comprehension group for the collection of supervised datasets and the evaluation of audio-based MLLMs.	Audio-related datasets: A-C-1, 2, 3, 4, 5, 6, 7, 8, 9	Specialists supporting A-C-1, 2, 3, 4, 5, 6, 7, 8, 9 skills, and MLLMs including Qwen-Audio-Chat, Qwen2-Audio-Instruct, Vitron-V1, GAMA, Pengi, WavLLM, SALMONN-7B SALMONN-13B SpeechGPT-7B-com, AudioGPT-GPT4, AnyGPT, PandaGPT-13B, ImageBind-LLM, ModaVerse-7b-v0, Unified-io-2-XXL, NExT-GPT-V1.5
8	Yaoting Wang	Working for Image group	Implement image-related MLLMs for evaluation	/	MLLMs including Qwen2-VL-7B, Qwen-VL-Chat, MoE-LLAVA-Phi2-2.7B-4e-384, mPLUG-Owl2-LLaMA2-7b, Phi-3.5-Vision-Instruct, Cambrian-1-8B, MiniGPT4-LLaMA2-7B
9	Junbao Zhou	Working for 3D group	Collect 3D-related datasets and Implement 3D-related Specialists and MLLMs for evaluation	3D-related datasets: D-C-1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	Specialists supporting D-C-1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 skills, and MLLMs including 3D-LLM-2.1B, PointLLM-7B, PointLLM-13B, 3D-VisTA
10	Jiahao Meng	Working for Video group	Collect video-related datasets and Implement video-related Specialists and MLLMs for evaluation	Video-related datasets: V-C-2, 3, 4, 20	Specialists supporting V-C-2, 3, 4, 20 skills, and MLLMs including Long-LLaVA-9B, DeepSeek-VL-2-small, DeepSeek-VL-2, LLaVA-One-Vision-7B, LLaVA-One-Vision-72B
11	Qingyu Shi	Working for Video group	Collect video-related datasets and Implement video-related Specialists and MLLMs for evaluation	Video-related datasets: V-G-1, 2, 3, 4	Specialists supporting V-G-1, 2, 3, 4 tasks, and MLLMs including VidAgent
12	Zhiyuan Zhou	Working for Image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	Image-related datasets: I-C-5, 7, 14, 15, 26, 28, 30, 34	Specialists supporting I-C-5, 7, 14, 15, 26, 28, 30, 34 skills, and MLLMs including InternVL2_5-2B, InternVL2_5-4B, InternVL2_5-8B, Monkey-10B-chat, DeepSeek-VL-7B-Chat
13	Liangtao Shi	Working for Image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	Image-related datasets: I-C-5, 13, 19, 26, 27, 28	Specialists supporting I-C-5, 13, 19, 26, 27, 28 skills, and MLLMs including InternVL-Chat-V1-5, Mini-InternVL-Chat-4B-V1-5, InternLM-XComposer2-VL-1.8B, GPT4RoI-7B, GLaMM

#	Name	Group	Role	Responsible Datasets	Responsible Models
14	Minghe Gao	Working for Image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	Image-related datasets: I-C-8, 9, 17, 21, 25, 26, 28, 31, 35	Specialists supporting I-C-8, 9, 17, 21, 25, 26, 28, 31, 35 skills, and MLLMs including BLIP2, miniMonkey, DeepSeek-VL-7B-Base, LISA
15	Daoan Zhang	Working for 3D group	Collect 3D-related datasets and Implement 3D-related Specialists and MLLMs for evaluation	3D-related datasets: D-G-1, 2, 3, 4, 5, 6, 7, 8, 9	Specialists supporting D-G-1, 2, 3, 4, 5, 6, 7, 8, 9 skills, and MLLMs including MotionGPT-T5, MotionGPT-LLaMA, AvatarGPT, LLaMA-mesh
16	Zhiqi Ge	Working for Image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	Image-related datasets: I-C-4, 26, 34	Specialists supporting I-C-4, 26, 34 skills, and MLLMs including Claude-3.5-Sonnet, Claude-3.5-Opus, Emu2-37B, DetGPT
17	Weiming Wu	Working for Image group	Implement image-related MLLMs for evaluation	/	MLLMs including Otter, Show-o, NExT-Chat, Yi-vision-v2, Qwen2-VL-72B
18	Siliang Tang	Working for Image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	Image-related datasets: I-C-26, 34	Specialists supporting I-C-26, 34 skills, and MLLMs including Claude-3.5-Sonnet, Claude-3.5-Opus, Emu2-37B, DetGPT
19	Kaihang Pan	Working for Image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	Image-related datasets: I-C-17, 22, 23, 38, 39	Specialists supporting I-C-17, 22, 23, 38, 39 skills, and MLLMs including Pixtral-12B, SEED-LLaMA-13B, LLaVA-NeXT-13B, LLaVA-NeXT-34B
20	Yaobo Ye	Working for Image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	Image-related datasets: I-C-3, 7, 28, I-G-1, 3, 4, 5, 6, 7, 8, 12, 14, 15	Specialists supporting I-C-3, 7, 28, I-G-1, 3, 4, 5, 6, 7, 8, 12, 14, 15 skills, and MLLMs including BLIP-3 (XGen-MM), CogVLM-Chat, ShareGPT4V-7B, ShareGPT4V-13B
21	Haobo Yuan	Working for Video group	Collect video-related datasets and Implement video-related Specialists and MLLMs for evaluation	Video-related datasets: V-C-5, 6, 7, 8, 9, 10, 11, 12, 13	Specialists supporting V-C-5, 6, 7, 8, 9, 10, 11, 12, 13 skills, and MLLMs including InternVL-2-8B, InternVL-2.5-8B, InternVL-2-26B, InternVL-2.5-26B
22	Tao Zhang	Working for Video group	Collect video-related datasets and Implement video-related Specialists and MLLMs for evaluation	Video-related datasets: V-C-14, 15, 16, 17, 18, 19	Specialists supporting V-C-14, 15, 16, 17, 18, 19 skills, and MLLMs including CoLVA-2B, CoLVA-4B, Sa2VA-8B, Sa2VA-26B
23	Tianjie Ju	Working for Language group	Collect language-related datasets and Implement language-related Specialists and MLLMs for evaluation	Language-related datasets: L-1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	Specialists supporting L-1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 skills, and MLLMs including Meta-Llama-3.1-8B-Instruct, Gemma-2-9b-it, GPT-J ChatGLM-6B, InternLM2-Chat-7B, Yi-lightning

#	Name	Group	Role	Responsible Datasets	Responsible Models
24	Zixiang Meng	Working for Image and Video group	Collect image, video-related datasets and Implement image, video-related Specialists and MLLMs for evaluation	Image-related datasets: I-C-6, 28, 38, and Video-related datasets: V-C-20	Specialists supporting I-C-6, 28, 38 and V-C-20 skills, and MLLMs including Gemini-1.5-Pro, Gemini-1.5-Flash, OMG-LLaVA-InternLM20B, Idefics3-8B-Llama3
25	Shilin Xu	Working for Video group	Collect video-related datasets and Implement video-related Specialists and MLLMs for evaluation	Video-related datasets: V-C-1, 2, 4	Specialists supporting V-C-1, 2, 4 skills, and MLLMs including InternVL-2.5-8B, InternVL-2.5-26B, Qwen2-VL-7B, Qwen2-VL-72B
26	Liyu Jia	Work for image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	image-related dataset collection: I-C-6, 17, and	MLLMs including GPT4-o-4096.
27	Wentao Hu	Work for image group	Collect image-related datasets and Implement image-related Specialists and MLLMs for evaluation	image-related dataset collection: I-C-17, 28	MLLMs including gpt-3.5-turbo, chatgpt4-o-latest.
28	Meng Luo	Working for Video group	Collect video-related datasets and Implement video-related Specialists and MLLMs for evaluation	Video-related datasets: V-C-2, V-G-1, 5, 6	Specialists supporting V-C-2, V-G-1, 5, 6 skills, and MLLMs including LM4LV
29	Jiebo Luo	Discussion & Advisory	Discussed the high-level directions and goals of the project. Provided important and insightful feedback for the overall system design.	/	/
30	Tat-Seng Chua	Discussion & Advisory	Discussed the high-level directions and goals of the project. Provided important and insightful feedback for the overall system design.	/	/
31	Hanwang Zhang	Project Supervision	1) Project co-supervisor, conceptualized the idea of General-Level, and the entire process. 2) Provided computing resources.	/	/
32	Shuicheng Yan	Project Supervision	1) Project co-supervisor, co-conceptualized the idea of General-Level, and supervised the entire process. 2) Provided computing resources.	/	/

Table 93: List of some contributors without authorship.

#	Name	Contribution
1	Zhengzhe Liu	Contributed to image group: assisted in image-oriented dataset preparation and model testing.
2	Zhongze Luo	Evaluating Yi-vision-v2 on 77 dataset; evaluating Show-o on 65 datasets.
3	Chunhan Li	Evaluating Show-o on 136 image comprehension datasets and partial 67 datasets.
4	Qirui Huang	Evaluating Yi-lightning on 117 NLP datasets and evaluating Show-o on 44 image generation.
5	Jiaxin Zhu	Assist in evaluating certain MLLMs on certain datasets.
6	Ming Lei	Evaluating Otter on certain datasets.
7	Zhangyu Wang	Evaluating Otter on certain datasets.
8	Lin Liu	Contributed to the preparation of image-related task data during phases 1 and 2.
9	Chengjie Zhou	Contributed to the preparation of NLP task data during phase 1.
10	Yucheng Han	Contributed to the preparation of image-related task data during phase 1.
11	Peng Zhou	Contributed to the preparation of image-related task data during phase 1.
12	Luanyuan Dai	Contributed to the preparation of image-related task data during phase 2.
13	Yuxuan Liu	Contributed to the preparation of image-related task data during phase 2.
14	Xun Jiang	Contributed to the preparation of image-related task data during phase 2.
15	Peisuo Li	Contributed to the preparation of image-related task data during phase 2.
16	Xu Zhang	Contributed to the preparation of image-related task data during phase 2.
17	Wenjie Zhuo	Contributed to the preparation of image-related task data during phase 2.
18	Lianyuan Fan	Contributed to 3D generation-related data collection during phase 2 (incomplete).