A BENCHMARK CONSTRUCTION

In this section, we introduce the collected datasets and the corresponding re-formulation procedures in detail. The statistics of the re-formulated datasets are provided in Table 7 and Table 8.

A.1 Coarse-Grained Perception

For the Flowers102 dataset, we employ the complete validation set for evaluation purposes. However, for CIFAR10, ImageNet-1K, Pets37, and VizWiz, we perform random subsampling of 10%. Concerning the TDIUC dataset, given that certain models in their training phase utilized a portion of the TDIUC dataset originating from the Visual Genome, we initially exclude this subset of data to prevent potential data leakage. Subsequently, we apply a shuffling operation to the entire TDIUC dataset and perform equidistant sampling, resulting in the selection of 2.5% of the sport_recognition data ($\text{TDIUC}_{\text{sport}}$) and 1% of the scene_recognition data ($\text{TDIUC}_{\text{scene}}$). In the case of MEDIC[3], we sample an equal number of samples from each label to balance the answer distribution.

For Flowers102 and Pets37, we randomly select three incorrect class labels, in addition to the correct label, from their original set of categories to form multiple-choice question options. For the TDIUC, we aggregate all answers for the same task to create an answer pool, and then utilize the same approach above to construct four answer options for multiple-choice questions.

For ImageNet-1K, we calculate similarities within its own set of 1000 categories using WordNet and selected the four options with the highest similarity to the correct class as choices (the highest one must be the right answer).

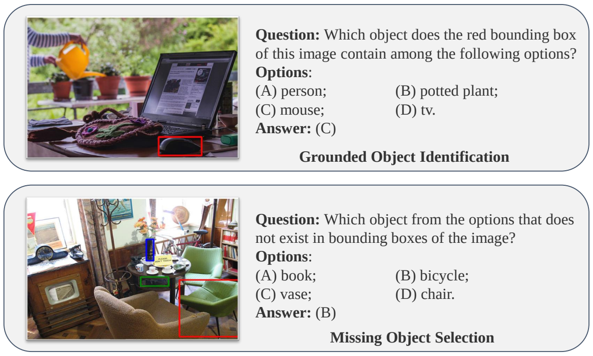
For CIFAR10, we initially employ WordNet to identify synonyms of the answers that are semantically related but not synonymous. These synonyms are then ranked based on their similarity. Subsequently, we manually adjust some of the less common candidate options. Finally, we likewise select the top four options with the highest similarity as all choices.

As for VizWiz, we re-formulate it into two benchmarks: VizWiz_2 as a binary classification task to determine whether there is any quality issue with the image. VizWiz_4 as a 4-choice question, requiring the model to determine the exact reason for the quality issue. We sort the issues related to image quality based on the number of votes in the annotations, the top one is considered the true label while the second to fourth options serve as negative choices.

For MEDIC [3], it is re-formulated to $\text{MEDIC}_{\text{dts}}$, a benchmark for disaster type selection (dts), we directly use all seven classification labels as choice options.

A.2 Fine-Grained Perception

For TDIUC [31], we initially exclude the subset sourced from Visual Genome [35] to prevent evaluation on the training data. Then, we shuffle the entire dataset and conducted an equidistant sampling strategy for task sample balance. Specifically, we sample 1% of the data for color ($\text{TDIUC}_{\text{color}}$), detection($\text{TDIUC}_{\text{detection}}$), and counting tasks ($\text{TDIUC}_{\text{counting}}$), and 2.5% for position tasks. As for the utility task ($\text{TDIUC}_{\text{utility}}$), we retain and utilized all 171 data samples. For answer options, we uniformly count all answers within the data and randomly selected three options other than the correct answer to form all four choices.



Question: Which object does the red bounding box of this image contain among the following options?
Options:
(A) person; (B) potted plant;
(C) mouse; (D) tv.
Answer: (C)

Grounded Object Identification

Question: Which object from the options that does not exist in bounding boxes of the image?
Options:
(A) book; (B) bicycle;
(C) vase; (D) chair.
Answer: (B)

Missing Object Selection

Figure 10: Examples of grounded fine-grained tasks.

Regarding RefCOCO [92], we re-formulate the referring expression selection ($\text{RefCOCO}_{\text{res}}$) task, in which the LVLMs are supposed to select the correct referring expression from the options based on the image region in the bounding box. We sample an equal number of samples from each object category, in order to balance the correct referring expression categories appearing in the questions. As for negative options in each question, we sample the negative referring expression from a distinct subcategory within the same category as the positive sample.

For MSCOCO [45], we re-formulate four tasks as follows: object counting (counting), multiple class identification ($\text{MSCOCO}_{\text{mci}}$), grounded object identification ($\text{MSCOCO}_{\text{goi}}$) and missing object selection ($\text{MSCOCO}_{\text{mos}}$) for object-level evaluation. The multiple class identification task aims to evaluate the LVLM's ability of object classification. Further, the grounded object identification and missing object selection tasks concentrate on object perception within a specified region of interest. The former allows models to assess which object exists within the given bounding box of the image, while the latter asks models to judge which object disappears within all the given bounding boxes of the image.

For the multiple class identification and grounded object identification tasks, we randomly sample 300 object annotations from each super-category in the valid split to ensure balance. This results in a total of 3600 evaluation data samples for each task. For the mos task, we filter out the objects with the height and width of their bounding boxes smaller than 50 and finally get 2479 samples. As for options generation, we employ a hierarchical strategy. For the multiple class identification task, we begin by randomly selecting the object class from within the super-category of the target object. If there are insufficient options, we broaden our selection to all object categories. In tasks related to region, our initial step is to randomly choose object categories present in the image but do not meet the requirement specified in the question. In cases where this is not possible, we follow the sampling procedure used in the multiple-class identification task. The examples of these grounded fine-grained tasks as shown in Table 10. The counting task has the same setting as the counting task in the TDIUC dataset.

A.3 Scene Text Perception

For OCR, we use 6 original OCR benchmarks (including CUTE80 [65], IC15 [32], IIIT5K [57], COCO-Text [57], WordArt [85] and TextOCR [71]) as the evaluation tasks. Current OCR benchmarks utilize cropped images containing only target text as visual input

Task Name	Dataset Name	Data Source	Datset Split	# of Images	# of Samples
Coarse-grained Perception	Flowers102	Flowers102	val	818	818
	CIFAR10	CIFAR10	test	10000	10000
	ImageNet-1K	ImageNet-1K	val	50000	50000
	Pets37	Pets37	test	3669	3669
	VizWiz$_2$	VizWiz	val	4049	4049
	VizWiz$_4$	VizWiz	val	2167	2167
	TDIUC$_{sport}$	TDIUC	val	6001	8696
	TDIUC$_{scene}$	TDIUC	val	9219	21320
	MEDIC$_{dts}$	MEDIC	test	15688	15688
Fine-grained Perception	MSCOCO$_{mci}$	MSCOCO	val2017	2323	3600
	MSCOCO$_{goi}$	MSCOCO	val2017	2404	3600
	MSCOCO$_{mos}$	MSCOCO	val2017	2479	2479
	TDIUC$_{color}$	TDIUC	val	18808	38267
	TDIUC$_{utility}$	TDIUC	val	162	171
	TDIUC$_{postiion}$	TDIUC	val	7131	9247
	TDIUC$_{detection}$	TDIUC	val	21845	29122
	TDIUC$_{counting}$	TDIUC	val	26166	41991
	RefCOCO$_{res}$	RefCOCO	val	9397	34540
	MSCOCO$_{count}$	MSCOCO	val2014	513	513
Scene Text Perception	CUTE80	CUTE80	all	288	288
	IC15	IC15	test	1811	1811
	IIIT5K	IIIT5K	test	3000	3000
	COCO-Text	COCO-Text	val	9896	9896
	WordArt	WordArt	test	1511	1511
	TextOCR	TextOCR	val	3000	3000
	Grounded IC15	IC15	val	221	849
	Grounded COCO-Text	COCO-Text	val	1574	3000
	Grounded TextOCR	TextOCR	val	254	3000
	FUNSD	FUNSD	test	47	588
	POIE	POIE	test	750	6321
	SROIE	SROIE	test	347	1388
	TextVQA	TextVQA	val	3023	4508
	DocVQA	DocVQA	val	1286	5312
	OCR-VQA	OCR-VQA	test	3768	3944

Table 7: Dataset statistics of visual perception tasks in ReForm-Eval.

sources [52, 86]. To further assess text identification in complex visual contexts, we propose grounded OCR tasks (including gIC15, gCOCO-Text, and gTextOCR). Specifically, we filter out the bounding boxes containing target texts larger than 40x40 for better evaluation. The image, along with the bounding box annotations and the corresponding instruction, will be fed into the model for evaluation, which is similar to the grounded fine-grained tasks (i.e. MSCOCO$_{goi}$). For KIE, we utilize the test splits of 3 benchmarks (including SROIE [27], POIE [37] and FUNSD [29]) as the evaluation tasks. And for OCR-based VQA, we use 3 benchmarks (including TextVQA [70], DocVQA [56] and OCR-VQA [58]) as the evaluation tasks. We filter out the question-answer pairs that need to be inferred based on the scene texts.

A.4 Visually Grounded Reasoning

For VQAv2 [21], we sample 10% for reformulation owing to the extremely large population. Besides, since ViQuAE [38] provides relevant knowledge information for each question, we additionally construct K-ViQuAE with knowledge as context, which assesses models' reasoning ability hierarchically with ViQuAE [38]. For ScienceQA [54], only 2017 questions of all the 4241 test set are paired with an image, which are selected in our benchmark. Besides, original A-OKVQA [67] gives rationales for answering each question, therefore we construct A-OKVQRA and A-OKVQAR for hierarchical evaluation.

For VQAv2 [21], GQA [28], OK-VQA [55], VizWiz [22], ViQuAE [38] and Whoops [6], ChatGPT is employed to generate appropriate negative options, and the prompt template for querying is:

Task Name	Dataset Name	Data Source	Datset Split	# of Images	# of Samples
Spatial Understanding	CLEVR	CLEVR	val	5726	6900
	VSR	VSR	test	1074	1811
	MP3D-Spatial	MP3D	-	3341	4735
Cross-Modal Inference	COCO_{itm}	MSCOCO caption	val2017	5000	25014
	COCO_{its}	MSCOCO caption	val2017	5000	25014
	WikiHow	WikiHow	val	32194	32194
	Winoground	Winoground	all	800	800
	SNLI-VE	SNLI-VE	test	1000	17901
	MOCHEG	MOCHEG	test	1452	3385
Visually Grounded Reasoning	VQA v2	VQA v2	val2014	15638	21441
	GQA	GQA	testdev	398	12578
	Whoops	Whoops	all	498	3362
	OK-VQA	OK-VQA	val	5032	5045
	ScienceQA	ScienceQA	test	2017	2017
	VizWiz	VizWiz	val	4319	4319
	ViQuAE	ViQuAE	test	1105	1257
	K-ViQuAE	ViQuAE	test	1094	1245
	A-OKVQA	A-OKVQA	val	1122	1145
	A-OKVQRA	A-OKVQA	val	1122	1145
	A-OKVQAR	A-OKVQA	val	1122	1145
	ImageNetVC	ImageNetVC	all	3916	4076
Multi-Turn Dialogue	VQA-MT	VQA v2	val2014	1073	1073
	VisDial	VisDial	val2018	2064	2064
Visual Description	COCO	MSCOCO caption	val2017	5000	5000
	TextCaps	TextCaps	val	3166	3166
	NoCaps	NoCaps	val	4500	4500
	Flickr30K	Flickr30K	test	1000	1000

Table 8: Dataset statistics of visual cognition tasks in ReForm-Eval.

> You are a multiple-choice generator. Given a question and an answer, you need to generate three additional incorrect options while ensuring their plausibility and confusion.
> Question: {question}
> Answer: {correct answer}

Note that for yes or no questions, the negative option is directly derived as no or yes, and ChatGPT is not employed.

While ImageNetVC [83] randomly selects 3 candidate options from the correspondent answer set with the commonsense type of each question. As for ScienceQA [54] and A-OKVQA [67], we adopt their original options.

As for A-OKVQAR, the prompt template for querying ChatGPT to generate negative rationales is:

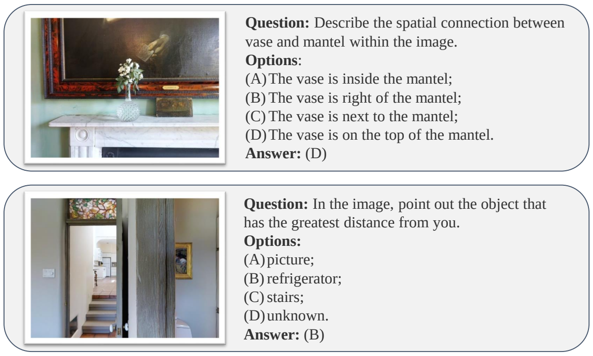
> You are a multiple-choice generator. Given a question and an answer, along with a rationale for that answer, you need to generate 3 counterfactual rationales. These counterfactual rationales should be contextually relevant while also being sufficiently distinct from the correct rationale.
> Question: {question}
> Answer: {correct answer}
> Rationale: {rationale}

A.5 Spatial Understanding

For CLEVR [30], we filter out the question types that do not involve spatial relations and randomly select 300 samples from each question type related to spatial relations. For different question types, we randomly select false options from their corresponding answer sets. In cases where some question types have insufficient options, we add “Not sure” and “Unknown” as false options to maintain the number of four options.

For VSR [46], the original dataset comprises captions that describe true or false spatial relations among objects in the corresponding images. We select image-caption pairs from the test split where the spatial descriptions are right and use them for our evaluation tasks. The false options are generated by randomly sampling different spatial relations from the test split.

MP3D [7] also known as Matterport3D, comprises a large-scale collection of RGB-D images captured from nearly 10,800 real indoor viewpoints with 50,811 object instance annotations. Based on this dataset, we extract two types of spatial relations to re-formulate our benchmark MP3D-Spatial: object-object level relations (left, right, above, under, on top of, and next to) and object-observer level relations (far and close). For spatial relations “on top of” and “next to”, we use 1,135 annotated samples for our task. For other relations, we utilize both bounding box information and depth to determine



Question: Describe the spatial connection between vase and mantel within the image.
Options:
(A) The vase is inside the mantel;
(B) The vase is right of the mantel;
(C) The vase is next to the mantel;
(D) The vase is on the top of the mantel.
Answer: (D)

Question: In the image, point out the object that has the greatest distance from you.
Options:
(A) picture;
(B) refrigerator;
(C) stairs;
(D) unknown.
Answer: (B)

Figure 11: Examples of spatial relation judgment in MP3D-Spatial.

the spatial relationships and extract 600 samples for each type of relation. As for false options, we randomly select non-matching spatial relations to serve as the incorrect options for our reformulated task. Figure 11 shows 2 re-formulated samples.

A.6 Cross-Modal Inference

In this section, we consider two kinds of tasks, including image text matching and visual entailment.

For MSCOCO [45], we re-formulate two tasks, including COCO image text matching (COCO$_{\text{itm}}$) and COCO image text selection (COCO$_{\text{its}}$). The matching task requests LVLMs to determine whether the given image and text are matched. The selection task instructs LVLMs to select the best-matched caption for the given image. We randomly sample image and text pairs as positive samples. For each image, we first find the negative images that have distinct but similar object categories. For each negative image, we find the most similar captions with the positive caption according to the object appearing in the sentence.

WikiHow [34] provides introductions to common skills. Within each skill, there are multiple crucial tasks, and each task is composed of several steps. We re-formulate the Wikihow image text selection task, in which given the task name, the LVLMs are supposed to choose the matching step description. We randomly sample visual and textual descriptions of the steps to form multiple-choice questions. To mine the hard negative options, we try to randomly take three samples from the same task as the positive sample. If the negative samples from the task level are insufficient, we then select some from the skill level and dataset level in turn.

For Winoground [75], we re-formulate a caption selection task, requiring models to choose the correct caption. Winoground provides captions for each image pair with identical words but in varying order, serving as options for the task.

For SNLI-VE [84], we re-formulate the visual entailment task, in which the LVLMs are required to determine whether the text can be inferred based on the image clues and should give answer of uncertainty when the evidence is insufficient. The options of multiple-choice question comprise “yes”, “not sure” and “no”. To balance the correct answer distribution, for each image, we sample an equal number of samples from each label.

For MOCHEG [88], we re-formulate the visual and textual entailment task, in which the LVLMs are supposed to determine whether the claim can be infered based on the visual and textual evidence and judge out whether the evidences are insufficient. The options consist of “supported”, “refuted” and “not enough information”.

A.7 Visual Description

We re-formulate the image captioning task from four dataset including MSCOCO [45], TextCaps [69], NoCaps [2] and Flickr30K [91]. In this task, the LVLMs are expected to generate a brief description for given image. Among these dataset, TextCaps additionally examines the optical character recognition capability of LVLMs by requesting models to pointing out the text in the image. We randomly sample these datasets for evaluation.

A.8 Multi-Turn Dialogue

To simulate a naive setup, we construct VQA-MT (VQA Multi-Turn) by considering multiple questions for the same image and gathering them into a multi-turn conversation. For VQA-MT, different images are accompanied by different amounts of questions in the re-formulated VQA v2 [21], only the images with more than 2 questions are kept. For images with more than 10 questions, only the first 10 questions are kept. All questions for the same image are arranged into a dialogue without inter-round dependencies. In the filtered dataset, there are 1073 image-dialogue pairs. The negative options are directly adopted from the re-formulated VQA v2.

As for VisDial [15], there is a 10-turn QA dialogue for each image. the original datasets provide 100 options for each question while. The prompt template for querying GPT-3.5 to generate negative options is:

> I will provide a question with the correct answer, please give me 3 incorrect options to help me get a single-choice question.
> Question: {question}
> Answer: {correct answer}

Different from the original VisDial to perform offline dialogue (the history contains correct answers), we perform online dialogue (the history contains the previous output of the models). To further investigate whether the performance of LVLMs changes with an increasing number of dialogue turns, we calculate the correlation coefficient between the accuracy and the number of dialogue turns.

B EVALUATION DETAILS

B.1 Implementation Details

Our benchmark and the evaluation framework are PyTorch-based. All experiments are conducted on 8 Tesla V100 GPUs. During the evaluation, half precision is used to accelerate the process.

To ensure fair comparisons between LVLMs, we try our best to keep the parameter setting aligned with the demo code provided by the original codebase. However, we limit the maximum number of tokens a model can generate for all LVLMs. It is set to 30 for most questions except the image-caption task where it is set to the upper quartile (the 75th percentile) of the reference caption length in the corresponding datasets. All input texts are formulated into conversations as required by different models, using the same system messages, roles, and separators. For BLIP-2, InstructBLIP, and Monkey, which have not been trained on multi-turn dialog, we use “Question” and “Answer” for the prompting format. As for the image input, we only consider single-image inputs, we use the same preprocess method mentioned in the original paper.

It is worth noting that ReForm-Eval comprises a total of over 500,000 evaluation instances across over 300,000 images, and considering the need for multiple tests for each instance, this results in significant computational cost. To this end, we further construct a subset by sub-sampling 10% data from the whole ReForm-Eval. All experiments conducted in this paper are based on the subset. We will open-source both the subset we use and the complete data for the research community. In Appendix C.4, we clarify the sampling method used and validate its effectiveness in making the subset balanced and consistent with the complete benchmark.

To avoid data leakage, we demonstrate the held-out sub-benchmark for fair comparison, namely ReForm-Eval-Sub, as the result of main experiments. 40 held-out datasets corresponding to all dimensions are listed below. In Coarse-grained Perception, all datasets are incorporated, a total of 9 datasets, including Flowers102, CIFAR10, ImageNet-1K, Pets37, VizWiz$_2$, VizWiz$_4$, TDIUC$_{\text{sport}}$, TDIUC$_{\text{scene}}$ and MEDIC$_{\text{dts}}$; Fine-grained Perception involves 9 datasets: MSCOCO$_{\text{count}}$, MSCOCO$_{\text{mci}}$, MSCOCO$_{\text{goi}}$, MSCOCO$_{\text{mos}}$, TDIUC$_{\text{color}}$, TDIUC$_{\text{utility}}$, TDIUC$_{\text{position}}$, TDIUC$_{\text{detection}}$ and TDIUC$_{\text{counting}}$; Scene Text Perception includes 9 datasets, including CUTE80, IC15, IIIT5K, WordArt, Ground IC15, FUNSD, POIE, SROIE and DocVQA; Visually Grounded Reasoning includes 4 datasets, which are Whoops, ViQuAE, K-ViQuAE and ImageNetVC; All datasets remain in Spatial Understanding, including 3 datasets: CLEVR, VSR and MP3D-Spatial; Cross-modal Inference involves 4 datasets, including WikiHow, Winoground, SNLI-VE and MOCHEG; Visual Description involves only 1 dataset of NoCaps; Multi-Turn Dialogue also involves 1 dataset of VisDial.

Please note that instability evaluation require multiple tests for the same samples, which introduce excessive cost for API-based proprietary models, so we do not include the results of instability for these models.

In Section 6.3 and Section 6.5. The results are obtained by excluding ~13B and proprietary models because we aim to eliminate the influence of model size and prevent the information in the images from becoming overly complex.

B.2 Models

In this section, we introduce the evaluated LVLMs in detail. For each method, we identify the version assessed in this paper if multiple variants are provided by the method. Additionally, we summarize the architecture of LVLMs in Table 9.

BLIP-2. BLIP-2 [40] is pre-trained in two stages: the representation learning stage and the generative learning stage, where the image encoder and the LLM are frozen and only a lightweight Q-Former is trained for bridging the modality gap. "blip2-pretrain-flant5xl" is evaluated in our experiment.

InstructBLIP. InstructBLIP [14] further extends BLIP-2 with task-oriented instruct tuning, pre-trained with Vicuna using the same procedure as BLIP-2. Additionally, an instruction-aware Q-Former module is proposed in InsturctBLIP, which takes in the instruction text tokens as additional input to the Q-Former. During instruction tuning, only parameters of Q-Former are fine-tuned based on pre-trained checkpoints, while keeping both the image encoder and the LLM frozen. We take "blip2-instruct-vicuna7b" and "blip2-instruct-flant5xl" as evaluation versions.

MiniGPT-4. MiniGPT4 [99] adds a trainable single projection layer based on BLIP-2 and also adopts a two-stage training approach, where the first stage is pre-training the model on large aligned image-text pairs and the second stage is instruction tuning with a smaller but high-quality image-text dataset with a designed conversational template. During training, the image encoder, the LLM, and the Q-Former are all frozen. "pretrained-minigpt4-7b" is used in our setup.

LLaVA-1.0. LLaVA-1.0 [50] employs a linear layer to convert visual features into the language embedding space, with a pre-training and instruction tuning stage. During pre-training, both the visual encoder and LLM weights were frozen. Then, keeping only the visual encoder weights frozen, the weights of the projection layer and LLM in LLaVA are updated with generated instruction data. In our experiment, "liuhaotian/LLaVA-7b-delta-v0" and "liuhaotian/llava-llama-2-7b-chat-lightning-lora-preview" are used for evaluation.

mPLUG-Owl. mPLUG-Owl [89] proposes a novel training paradigm with a two-stage fashion. During pre-training, mPLUG-Owl incorporates a trainable visual encoder and a visual abstractor, while maintaining the LLM frozen. In the stage of instruction tuning, language-only and multi-modal instruction data are used to fine-tune a LoRA module on the LLM. "MAGAer13/mplug-owl-llama-7b" is used in our experiment, but LoRA is not implemented in this version.

ImageBind-LLM. ImageBind-LLM [23] adopts a two-stage training pipeline. In the pre-training stage, a learnable bind network and a gating factor are updated. The bind network transforms image features, while the gating factor weights the transformed image features to determine the magnitude of injecting visual semantics and the result is added to each word token for each layer. In the instruction tuning stage, a mixture of language instruction data and visual instruction data is used to update partial parameters in LLaMA by LoRA and bias-norm tuning. We utilize "Cxxs/ImageBind-LLM/7B" for evaluation.

LLaMA-Adapter V2. In LLaMA-Adapter V2 [18], a joint training paradigm is proposed, where only the visual projection layers and early zero-initialized attention with gating are pre-trained using image-text data, while the late adaptation prompts with zero gating, the unfrozen norm, newly added bias, and scale factors are implemented for learning from the language-only instruction data. "LLaMA-Adapter-V2-BIAS-7B" is applied for evaluation.

Multimodal-GPT (mmGPT). Multimodal-GPT [20] is fine-tuned from OpenFlamingo, where the whole open-flamingo model is frozen and the LoRA module is added and updated to the self-attention, cross-attention, and FFN part in the LLM, using language-only and multimodal instruction data. "mmgpt-lora-v0-release" is used for evaluation. To simplify, we refer to it as "mmGPT".

PandaGPT. PandaGPT [73] utilizes a one-stage training method using a combination of 160k image-language instruction-following data from MiniGPT-4 and LLaVA, where only two components are trained: a linear projection matrix connecting the visual representation generated by ImageBind to Vicuna, and additional LoRA weights applied to Vicuna attention modules. "pandagpt-7b-max-len-1024" is evaluated as our implemented version.

Model	Model Architecture			
	Vis Encoder	LLM	Connection Module	#oP
BLIP-2	ViT-G/14	FlanT5-XL	Q-Former	3.9B
InstructBLIP$_F$	ViT-G/14	FlanT5-XL	Q-Former	4.0B
InstructBLIP$_V$	ViT-G/14	Vicuna-7B	Q-Former	7.9B
LLaVA-1.0-7B$_V$	ViT-L/14	Vicuna-7B	Linear	7.1B
LLaVA-1.0-7B$_{L_2}$	ViT-L/14	LLaMA2-7B	Linear	7.1B
MiniGPT4	ViT-G/14	Vicuna-7B	Q-Former+Linear	7.8B
mPLUG-Owl	ViT-L/14	LLaMA-7B	Perceiver	7.1B
PandaGPT	ImageBind	Vicuna-7B+LoRA	Linear	8.0B
ImageBindLLM	ImageBind	LLaMA-7B+LoRA+BT	BindNet+Gate	8.6B
LA-V2	ViT-L/14	LLaMA-7B+BT	Linear+Adapter+Gate	7.1B
mmGPT	ViT-L/14	LLaMA-7B+LoRA	Perceiver+Gate	8.4B
Shikra	ViT-L/14	Vicuna-7B	Linear	6.7B
Cheetor$_V$	ViT-G/14	Vicuna-7B	Query+Linear+Q-Former	7.8B
Cheetor$_{L_2}$	ViT-G/14	LLaMA2-Chat	Query+Linear+Q-Former	7.8B
BLIVA	ViT-G/14	Vicuna-7B	Q-Former+Linear	7.9B
LLaVA-1.5-7B$_V$	ViT-L/14	Vicuna-v1.5-7B	Linear	7.2B
MiniGPT-v2	ViT-G/14	LLaMA2-7B	Linear	8B
Qwen-VL-Chat	ViT-bigG/14	Qwen-7B	Resampler	9.7B
LLaVA-1.6-7B$_V$	ViT-L/14	Vicuna-v1.5-7B	Linear	7.1B
Monkey	ViT-bigG/14	Qwen-7B	Resampler	9.8B
Deepseek-VL	SAM-B + SigLIP-L	Deepseek-7B	MLP	7.3B
ShareGPT4V-7B	ViT-L/14	vicuna-v1.5-7B	Linear	7.2B
ShareGPT4V-13B	ViT-L/14	vicuna-v1.5-13B	Linear	13.4B
OmniLMM-12B*	Eva-02-5B	Zephyr-7B-β	Resampler	13B
LLaVA-1.5-13B$_V$	ViT-L/14	Vicuna-v1.5-13B	Linear	13.4B
LLaVA-1.6-13B$_V$	ViT-L/14	Vicuna-v1.5-13B	Linear	13.4B
Qwen-VL-Max	Unknown	Unknown	Unknown	Unknown
Gemini-1.0-Pro-Vis	Unknown	Unknown	Unknown	Unknown
GPT-4V	Unknown	Unknown	Unknown	Unknown

PS: Underlined represents a trainable component. The training detail of **OmniLMM-12B** is not fully disclosed.

Table 9: Model architecture of different LVLMs. "#oP"is the number of total parameters. "BT" represents bias-tuning. "BindNet" represents bind network. "Unknown" denotes the specific detail is unknown.

Shikra. Shikra [9] consists of a vision encoder, an alignment layer, and an LLM. This model is trained in two stages, where both the fully connected layer and the entire LLM are trained and the visual encoder is frozen. We select "shikras/shikra-7b-delta-v1" for our evaluation in this experiment.

Cheetor. Cheetor [41] is initialized from BLIP-2 and pre-trains Q-Former that matches Vicuna and LLaMA2. A lightweight CLORI module is introduced that leverages the sophisticated reasoning ability of LLMs to control the Q-Former to conditionally extract specific visual features, and further re-inject them into the LLM. During training, only a set of query embeddings and two linear projection layers need to be updated. "cheetah-llama2-7b" and "cheetah-vicuna-7b" are specifically used for assessment.

BLIVA. BLIVA [25] is initialized from a pre-trained InstructBLIP and merges learned query embeddings output by the Q-Former with projected encoded patch embeddings. Demonstrating a two-stage training paradigm, the patch embeddings projection layer is pre-trained and both the Q-Former and the project layer are fine-tuned by instruction tuning data. "mlpc-lab/BLIVA-Vicuna" is employed under evaluation in our experiment.

LLaVA-1.5. LLaVA-1.5 [48] increases the resolution of input images by using CLIP-ViT-L-336px as the vision encoder. Besides, it adopts a two-layer MLP projection to align the visual features to the word embedding space of LLM. Furthermore, LLaVA-1.5 adds some task-oriented instructing tuning data to improve its performance. We use "liuhaotian/llava-v1.5-7b" and "liuhaotian/llava-v1.5-13b" for evaluation.

MiniGPT-v2. MiniGPT-v2 [8] increases the resolution of input images to 448x448 and concatenate 4 adjacent visual tokens and project them together into one single embedding for training and inference efficiency. Unique identifiers for different tasks are also proposed to enable the model better distinguish each task instruction and improve the model learning efficiency for each task.

LLaVA-1.6. LLaVA-1.6 [49] supports three aspect ratios of input images, up to 672x672, 336x1344, 1344x336 resolution. High-quality instruction tuning data and multimodal document/chart data are introduced to enhance the visual reasoning ability. We employ "liuhaotian/llava-v1.6-vicuna-7b" and "liuhaotian/llava-v1.6-vicuna-13b" for assessment in our experiment.

ShareGPT4V. ShareGPT4V [10] follows the design of LLaVA-1.5, using a two-layer MLP as the projection layer. ShareGPT4V introduces a high quality image captions dataset generated by GPT4V and its own caption model, which is used to its pretraining and supervised finetuning stage. "ShareGPT4V-7B" and "ShareGPT4V-13B" are used for evaluation.

Qwen-VL-Chat. Qwen-VL-Chat [5] uses Qwen-7B as the LLM backbone and ViT-G as the initialization of the visual encoder. The resolution of input images is 448×448. These modules are connected by a randomly initialized cross-attention layer. Qwen-VL-Chat is the instruction-tuned vision-language chatbot based on Qwen-VL, which supports more flexible interaction.

Monkey. Monkey [44] enhances the LVLM architecture by dividing images into multiple patches of 448×448 and introducing several adapters for encoding the patches. As a result, Monkey is able to handle 1344×896 pixels and capture more details. Further more, Monkey propose a method to generate multi-level description for training the LVLM.

Deepseek-VL. Deepseek-VL [53] is a LVLM built on Deepseek. The training procedure is divided into 3 phases: VL adaptor pre-training, joint V-L pre-training, and instruct tuning. High-quality text-only data is introduced in these phases to enhance the text understanding ability. In this paper, we utilize the "Deepseek-VL-7B-chat" for evaluation.

OmniLMM. OmniLMM [1] is a LVLM where EVA-2-5B is connected with Zephyr-7B-β, a RLHF fine-tuned version of Mistral-7B. However, the training detail of OmniLMM is not fully disclosed. In this paper, we use "OmniLMM-12B" for evaluation.

Qwen-VL-Max. Qwen-VL-Max stands the most capable large vision language model of Qwen-VL model family by Alibaba [5]. Notably, it offers robust support for high-definition images surpassing one million pixels and those with extreme aspect ratios. In rigorous evaluations across various text-image multimodal challenges, it demonstrates performance parity with industry-leading models such as Gemini Ultra and GPT-4V. In this paper, we employ the latest version of Qwen-VL-Max, as introduced on January 25, 2024.

GPT-4V. GPT-4V is a multimodal version of the powerful GPT-4 model developed by OpenAI [61], which combines together the language comprehension and image processing capabilities. Although details are missed, the training process of GPT-4V was believed the same as that of GPT-4, but used a large number of text-image paired data from the Internet. This model not only recognizes objects in images, but also has the ability to understand image context, subtle differences, and nuances. In this paper, we employ the version of "gpt-4-turbo-2024-04-09".

Gemini-1.0-ProVis. Gemini-1.0-ProVis [74] is a multimodal large model launched by Google in December 2023. Developers can access this model for free in the development platform of Google AI Studio. Compared with GPT-4V, Gemini Pro surpassed it with a high score of 1933.4 in terms of comprehensive performance on the multimodal proprietary benchmark MME, demonstrating its comprehensive advantages in perception and cognition.

C COMPLEMENTARY RESULTS AND ANALYSIS

C.1 Per-Dimension Results and Analysis

In this section, we will provide the complete results and corresponding analysis of all capability dimensions. It is worth noting that the average performance listed in this section is calculated with results in held-out datasets. Unlike the setup in the main article, the **best results** and the runner-up in this section is marked across all models, without distinguishing between groups.

C.1.1 Results on Coarse-grained Perception. Tabel 15 and Table 16 provide results of coarse-grained perception tasks, including image classification and scene recognition. For image classification, api-based proprietary models demonstrate significant advantages in most tasks except CIFAR10. We speculate this is attributed to the low resolutions of images in CIFAR10. Monkey, Qwen-VL-Max, Gemini-1.0-Pro-Vis, GPT-4V fail on this task since they rely on high input resolutions and can not adapt to low-quality images well. Another finding is that image quality assessment in VizWiz is challenging to current LVLMs, implying these models can not fully understanding the attributes of the image even when they are good at understanding the contents in images. Under likelihood evaluation, the while-box evaluation method reveals the effectiveness of Qwen-VL-Chat, Qwen-VL-Chat even outperforms ~13B models in several tasks.

In terms of scene recognition tasks, the trend is similar to that in image classification tasks. BLIP-2 and InstructBLIP perform well on these tasks, indicating that Q-Former connection can well capture the global semantic in images.

In general, proprietary models are the best models in addressing coarse-grained perceptions tasks. Among open-source models, OmniLMM, LLaVA-1.6, and Qwen-VL-Chat demonstrate outstanding capabilities.

C.1.2 Results on Fine-grained Perception. Tabel 17 and Table 18 provide results of fine-grained perception tasks, including Object Perception and Object Grounding. For object perception, Qwen-VL-Max and OmniLLM-12B dominate most tasks, especially when evaluated with the generation evaluation strategy. Under likelihood evaluation, OmniLLM-12B, ShareGPT4V-13B and LLaVA-1.5-13B$_V$ are comparable. Considering the results of MSCOCO$_{goi}$, most models are able to solve a part of the questions, indicating that LVLMs are able to understand the bounding boxes in images and the grounded questions. Bounding box labeling can be an optional method to provide locality information without the need for understanding continuous input. As for object grounding, the task is quite difficult for most 7B models, while 13B models and proprietary API-based models achieve significantly good performance under generation evaluation. Nevertheless, all the 7B and 13B models struggle under

Dataset	Metric	Generation					Likelihood				
		1%	2%	10%	20%	100%	1%	2%	10%	20%	100%
VQAv2	ρ	0.9861	0.9948	0.9989	0.9996	1	0.9689	0.9857	0.9970	0.9991	1
	$\bar{d}$	3.15	1.71	0.56	0.37	0	2.65	2.12	0.87	0.50	0
Flowers102	ρ	0.9575	0.9559	0.9794	0.9336	1	0.7984	0.7861	0.9131	0.9727	1
	$\bar{d}$	8.57	9.03	4.38	1.70	0	12.18	10.69	3.25	2.93	0

Table 10: Evaluation results under different sampling ratios on the VQAv2 and Flowers102 benchmark. We derive the results of different models on test sub-benchmarks under each sampling ratio, and calculate the correlation coefficient ρ and average absolute deviation $\bar{d}$ of these results compared to the results on the complete test benchmark.

Dataset	Size	Generation					Likelihood				
		1%	2%	10%	20%	100%	1%	2%	10%	20%	100%
VQAv2	21441	5.22	2.90	0.44	0.19	0	8.63	5.06	0.81	0.26	0
Flowers102	818	169.79	85.60	18.83	6.62	0	243.01	174.59	17.79	10.66	0

Table 11: Variance of accuracy(%) under different sampling ratios. We repeatedly sample a certain percentage of the data for 10 times. We derive the accuracy each time and then compute the variance for each model. The final variance value is averaged across the models.

likelihood evaluation. We speculate that this is because there are only subtle differences between options in these questions, such as “the person on the left” and “the person on the right”. In generation evaluation, all options are provided in the context, helping the models with strong instruct understanding abilities to distinguish between them. As for likelihood evaluation, options are provided to the model separately, models may not be able to distinguish them effectively.

C.1.3 Results on Scene Text Perception.

Table 21 and Table 22 provide results of scene text perception, which consists of OCR, Grounded OCR, KIE and OCR-based VQA tasks. Since the scene text perception task requires the model output to contain the target tokens in the image, only generation evaluation is conducted. Qwen-VL-Max and GPT4-V perform the best in almost all tasks while Gemini-1.0-Pro-Vis also demonstrates its effectiveness. For open-source models, OmniLMM consistently dominates the OCR and GroundOCR tasks, while Monkey and Qwen-VL-Chat perform well in KIE and OCR-based VQA tasks. Notably, the trend suggests that larger models tend to outperform smaller ones, owing to their greater capacity for contextual modeling. Furthermore, training with OCR-related tasks, as exemplified by models from the Qwen-VL model family, notably enhances performance in scene text perception tasks. In general, proprietary models are dominate in the scene text perception tasks.

C.1.4 Results on Visually Grounded Reasoning.

Table 19 and Table 20 provide results of visually grounded reasoning, which consists of VQA and KVQA. For generation evaluation, Qwen-VL-Max achieves top-2 performance on nearly all the datasets for VGR tasks, and Gemini-1.0-Pro-Vis follows closely behind. Besides, OmniLMM-12B also exhibits excellent performance on GQA, VQA v2, Whoops, OK-VQA and ScienceQA. Surprisingly, GPT-4V doesn’t show leading performance on VGR tasks. While in likelihood evaluation, we can only evaluate those open source models. And we can find that some 7B models achieve awesome and competitive performance. On average, Monkey and Qwen-VL-Chat are the top-2 models.

We also conducted a hierarchical evaluation of LVLMs’ external knowledge incorporation and reasoning abilities. Comparing the results of ViQuAE and K-ViQuAE, as well as A-OKVQA and A-OKVQRA, it is evident that, with the provision of external knowledge, the performance of most models has significantly improved.

C.1.5 Results on Spatial Understanding.

Table 23 provides results of the spatial understanding capability dimension, which includes relation judgment and space-based perception tasks. For generation evaluation, OmniLMM and Qwen-VL-Max emerge as dominant contenders across the majority of tasks. Meanwhile, in likelihood evaluation, alongside OmniLMM, ShareGPT4V and LLaVA also assert their leadership positions.

In the context of the spatial relation judgment task, it’s noteworthy that performance on the MP3D-Spatial dataset appears relatively poorer compared to the other two datasets. This discrepancy is believed to stem from the fact that MP3D-Spatial is sampled from real-world navigation environments, inherently more intricate and potentially divergent from the training data of the LVLM. For space-based perception tasks, likelihood evaluation yields better results than generation evaluation, especially for LLaVA, mPLUG-Owl, LA-V2, MiniGPT-v2, Shikra and Cheetor. This might be attributed to the high demand for spatial reasoning skills for this task, thereby placing a greater emphasis on the image comprehension abilities of visual backbones. Most of these models use ViT-L, which lacks robust spatial semantic understanding.

C.1.6 Results on Cross-modal Inference.

Table 26 provides results of the cross-modal inference capability dimension, which includes image-text matching tasks and visual entailment tasks. For the image-text matching task in MSCOCO, we consider a one-to-one setup of the naive image-text matching and a one-to-four selection setup of image-text selection. Two FlanT5 based models of BLIP series and proprietary models perform well in both setups under the generation evaluation. However, the performance of most models has reduced under the likelihood evaluation for image-text selection, we attribute this to the same reason that is mentioned in the analysis of referring expression selection in Appendix C.1.2. Unlike MSCOCO, WikiHow considers the scenarios to match images and abstract instructions, while Winogroud uses negative options with only minor word-level modifications. These pose significant challenges for the models, resulting in a noticeable decrease in accuracy. However, proprietary models maintains a lead, followed by ~13B model. Regarding the visual entailment task, apart from the two models based on FlanT5 and proprietary models, the performance of the other models is not promising. In summary, we believe that current LVLMs still have relatively weak capabilities in logical reasoning and understanding fine-grained textual details.

C.1.7 Results on Visual Description. Table 24 and Table 25 provides image captioning results of the visual description capability dimension. We choose CIDEr metric to estimate visual description capability while providing BLEU-4, METEOR and ROUGE-L results for additional references. As mentioned in previous work [86], these datasets require concise captions while most LVLMs tend to generate detailed descriptions. Therefore, the performance of most models is not satisfying enough. For this task, PandaGPT always generates a sentence starting with "the image features" and MiniGPT-v2 is also accustomed to outputting long guiding phrases, leading to their limited performances. At the same time, ShareGPT4V-13B dominates the task because ShareGPT4V-13B is able to provide short captions. To adapt to the development of LVLMs, there is a strong need for a benchmark for evaluating detailed description capabilities.

C.1.8 Results on Multi-Turn Dialogue. Table 27 provides results of the multi-turn Dialogue task. Qwen-VL-Max and OmniLMM perform the best in this task while the effectiveness of Qwen-VL-Chat and ShareGPT4V is demonstrated under likelihood evaluation. In general, larger models perform better due to the larger capacity in modeling the context. It is worthy noting that for models like Monkey that has not been trained on multi-turn instruction data, the performance is not satisfactory as in single-turn questions. Multi-turn data should be incorporated during training to further improve existing LVLMs.

C.2 Effect of In-context Sample

Here we declare the setting in Section 6.2. The experiment is conducted on the re-formulated VQA v2.0, we sample a subset of 1000 samples for efficiency. The format hit (compliance) rate is the proportion of reponses in the desired format, namely a string containing options enclosed in parentheses. Another insight gleaned in our experiment is that the number of options in the in-context sample should not be fewer than the number of options in the target questions.

C.3 Annotation Cost

In this section, we will introduce the cost details of different annotation methods. We compare the average costs of using ChatGPT-3.5, GPT-4V and manual annotation for negative options. Firstly, we calculate the average number of words for questions, options and prompts, which is 69, and assuming that a word will be tokenized into 2 tokens on average. Therefore, the average number of input tokens is 138. With an average output of 3 options, it costs approximately $0.00035 per instance using ChatGPT-3.5 for annotation. And for GPT-4V, an input image with the lowest resolution contains 255 tokens, costing $0.00429 per instance, which is about 12 times the cost of ChatGPT-3.5. As for human annotations, some of our researchers attempted to annotate 100 samples and it spent approximately 80 minutes. Considering that crowdworkers might take roughly twice as long due to their unfamiliarity with the task setting, we estimate that manually annotation costs $0.126 per instance, which is almost 30 times the cost of GPT-4V. Above all, using ChatGPT-3.5 for annotation is both efficient and cost-effective.

C.4 Influence of Sampling Methods

Here we clarify the sampling method used in this paper. Since the Reform-Eval covers 61 benchmarks and contains more than 500,000 evaluation samples, we employ ***a balanced sampling strategy*** to ensure the sampled subsets maintain similar characteristics of original benchmarks, thereby enhancing the robustness of the evaluation. Within each benchmark, we perform a balanced sampling based on the distribution of the original data, at a rate of 10% except for three cases: (1) when the size of the original benchmark is less than 1000, we keep the whole benchmark for a stable evaluation; (2) when the original benchmark has more than 10,000 evaluation samples, we filter the data and then conduct the sampling process; (3) for benchmarks used in Multi-turn Dialogue dimension, we retain all evaluation samples as the total sample volume is moderate in this dimension (~3000). It is worth noting that our calculation method is to first compute the scores of the models on each evaluation benchmark and then average across the benchmarks to obtain the final score, rather than mixing all the evaluation benchmarks together and then computing the overall score on the mixed dataset. Such sampling methods guarantee that the results on each evaluation benchmark are stable and reliable, leading to relative fairness and balance across all benchmarks.

We further analyze whether the shrink or expansion of the dataset size will change the evaluation results. We conduct several experiments on the VQAv2 benchmark and Flowers102 benchmark (the evaluation sample size is 21441 and 818, respectively) in the Coarse-Grained Perception and Visually Grounded Reasoning dimensions. Table 10 demonstrates that the more data sampled, the better the stability of the results, and the more consistent they are with the evaluation on the complete dataset. A 10% sampling ratio can achieve a good balance between evaluation efficiency and consistency.

Moreover, for larger datasets, the sampling ratio has little impact on the results; for smaller datasets, the sampling ratio greatly affects the results (see Table 11). Therefore, we generally perform balanced distribution sampling for larger datasets and retain the entire dataset for smaller datasets.

Model	Generation			Likelihood
	Instruct	Option Order	Option Mark	Instruct
BLIP-2$_F$	0.029	0.276	0.107	0.037
InstructBLIP$_V$	0.038	0.414	0.182	0.018
LLaVA-1.0-7B$_V$	0.197	0.606	0.299	0.105
MiniGPT4	0.113	0.647	0.194	0.043
mPLUG-Owl	0.330	0.706	0.406	0.046
PandaGPT	0.125	0.592	0.198	0.117
ImageBindLLM	0.159	0.709	0.498	0.024
LA-V2	0.382	0.682	0.518	0.032
mmGPT	0.577	0.763	0.601	0.030
Shikra	0.028	0.617	0.206	0.054
Cheetor$_{L_2}$	0.051	0.476	0.163	0.058
BLIVA	0.128	0.610	0.204	0.023
LLaVA-1.5-7B$_V$	0.068	0.363	0.121	0.067
MiniGPT-v2	0.110	0.530	0.104	0.136
Qwen-VL-Chat	0.037	0.398	0.123	0.124
LLaVA-1.6-7B$_V$	0.081	0.311	0.143	0.068
Monkey	0.034	0.396	0.201	0.051
Deepseek-VL	0.025	0.224	0.077	0.070
ShareGPT4V-7B	0.040	0.327	0.090	0.061
~7B Avg.	**0.134**	**0.507**	**0.233**	**0.061**
ShareGPT4V-13B	0.073	0.292	0.127	0.054
OmniLMM-12B	0.034	0.204	0.043	0.101
LLaVA-1.5-13B$_V$	0.037	0.270	0.065	0.049
LLaVA-1.6-13B$_V$	0.039	0.281	0.072	0.041
~13B Avg.	**0.046**	**0.261**	**0.077**	**0.061**
Qwen-VL-Max	0.120	0.196	0.106	-
Gemini-1.0-ProV	0.047	0.17	0.043	-
GPT-4V	0.199	0.295	0.212	-
Pro. Avg.	**0.122**	**0.221**	**0.120**	-

Table 12: Instability of models caused by different random perturbations. "Pro." represents the proprietary group.

C.5 Instability

Table 12 provides the complete results of models' instability caused by different perturbations. Under the generation evaluation, all models are most sensitive to the order of options, followed by the option marks, and lastly, random instructions. FlanT5 models are the most stable models under the generation evaluation, showing that FlanT5 can well comprehend the multiple-choice questions. For likelihood evaluation, all models are stable since the evaluation strategy directly utilizes the characteristics of generative models.

To further perceive the influence of instruction perturbation on the answer accuracy, we analyze the above instruction perturbation results. As we employ different instructions to describe the same task, the accuracy of samples that follow each instruction can be calculated. For the accuracy of each instruction, we adopt the difference between the maximal and minimal accuracies to represent the model's instability level towards the instruction. The results are shown in Table 13. We discover that all models exhibit some fluctuations in accuracy, illustrating that LVLMs are sensitive to designed prompts. However, the fluctuations in accuracy under generation and likelihood evaluation of most LVLMs are both within an acceptable range. There are still models exhibiting fluctuations in accuracy exceeding 10%, indicating the restricted instruction-following capabilities of LVLMs. In general, LVLMs require further improvements to enhance its ability to understand and follow diverse instructions.

The above phenomenon indicates that single tests under specific prompts are unstable and may introduce bias, while ReForm-Eval comprehensively considers various factors and can provide stable evaluation results.

Model	Generation	Likelihood
BLIP-2$_F$	11.94	6.97
InstructBLIP$_F$	9.45	5.97
InstructBLIP$_V$	7.46	5.97
LLaVA-1.0-7B$_V$	5.47	6.47
LLaVA-1.0-7B$_{L_2}$	15.42	7.46
MiniGPT4	4.48	5.97
mPLUG-Owl	4.98	6.47
PandaGPT	5.97	3.98
ImageBindLLM	2.99	3.98
LA-V2	6.47	6.47
mmGPT	6.46	9.95
Shikra	13.93	5.97
Cheetor$_V$	9.95	2.99
Cheetor$_{L_2}$	10.95	9.95
BLIVA	2.49	7.46
LLaVA-1.5-7B$_V$	2.99	3.48
MiniGPT-v2	7.96	14.93
Qwen-VL-Chat	1.99	2.49
LLaVA-1.6-7B$_V$	6.47	1.49
Monkey	1.49	1
Deepseek-VL	4.48	14.43
ShareGPT4V-7B	2.49	2.49
ShareGPT4V-13B	2.49	1.99
OmniLMM-12B	1.49	13.43
LLaVA-1.5-13B$_V$	6.47	7.96
LLaVA-1.6-13B$_V$	10.95	11.94
Qwen-VL-Max	2.99	-
Gemini-1.0-ProV	4.98	-
GPT-4V	1.99	-
Average	**6.13**	**6.60**

Table 13: The difference between the maximum and minimum accuracies of all instruction groups.

C.6 Option Preference

Option preference is a phenomenon in our benchmark that when uncertain about the answer, LVLMs prefer a particular option regardless of options' content. We verify the option preference inside the LVLMs in Figure 12. It has been observed that ImageBind-LLM, Shikra and BLIVA exhibit a preference for option "A" when confronted with uncertainty. MiniGPT4, mPLUG-Owl, PandaGPT, LA-V2, mmGPT, MiniGPT-v2 and Deepseek-VL show a strong preference for option "B". Other LVLMs show no obvious preference in this task. It's worth noting that predicted choice distribution under the likelihood evaluation method has no preference, as all options are considered in an unordered state.

The phenomenon of option preference contributes to the instability from random option order but reduces that from random instruction and option mark (as mentioned in Section 6.4). Concretely, when LVLMs are uncertain about answers, they select the ceratin option repeatedly for the essentially identical questions. As the option contents have been shuffled in random option mark mode, the LVLMs are regarded as selecting distinct answers. Regarding random instruction and option mark situations, LVLMs are firm in their answers regardless variation of question form. This also highlights the importance of introducing randomness by shuffling the options in ReForm-Eval.

Model	**Generation**	**Likelihood**
BLIP-2$_F$	-0.44	0.13
InstructBLIP$_F$	-0.53	-0.03
InstructBLIP$_V$	-0.40	-0.27
LLaVA-1.0-7B$_V$	-0.51	-0.01
LLaVA-1.0-7B$_{L_2}$	-0.24	0.28
MiniGPT4	-0.28	0.05
mPLUG-Owl	-0.42	0.15
PandaGPT	-0.22	0.02
ImageBindLLM	-0.09	0.01
LA-V2	-0.43	0.22
mmGPT	-0.15	0.15
Shikra	-0.19	0.13
Cheetor$_V$	-0.48	0.14
Cheetor$_{L_2}$	-0.30	0.15
BLIVA	0.05	-0.26
LLaVA-1.5-7B$_V$	-0.42	0.03
MiniGPT-v2	-0.46	0.05
Qwen-VL-Chat	-0.52	-0.48
LLaVA-1.6-7B$_V$	-0.67	0.07
Monkey	-0.71	-0.20
Deepseek-VL	-0.54	-0.18
ShareGPT4V-7B	-0.74	-0.04
ShareGPT4V-13B	-0.63	-0.09
OmniLMM-12B	-0.60	-0.12
LLaVA-1.5-13B$_V$	-0.52	0.00
LLaVA-1.6-13B$_V$	-0.57	0.09
Average	**-0.42**	**0.00**

Table 14: The correlation between instability and accuracy across all open-source LVLMs.

C.7 Correlation Between Instability and Accuracy

We additionally calculate the correlation between instability and accuracy to delve into their relation, as shown in Table 14. The negative correlation between instability and accuracy under generation method is apparent, where the high instability reflects the reduced accuracy. The average correlation under likelihood method is zero, demonstrating low relativity between them here. As the randomness perturbations under likelihood method are rare, their instability is generally low, leading to the unrelated relation.

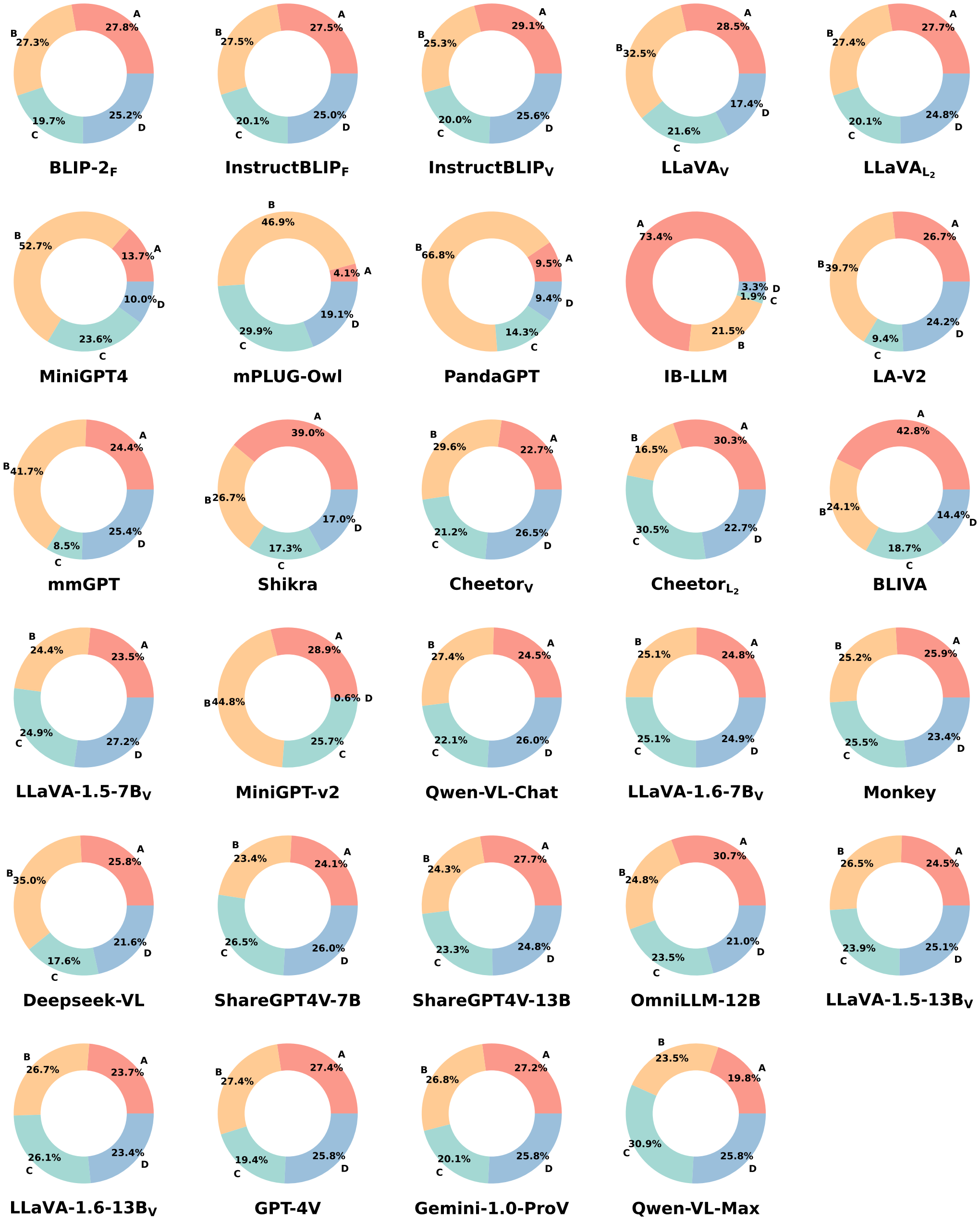


Figure 12: The choice distribution of prediction of all LVLMs in COCO image text selection task (ITS) under generation method. The ground truth choice distribution is uniform as we have shuffled the options.

Model	Image Classification											
	Flowers102		**CIFAR10**		**ImageNet-1K**		**Pets37**		**VizWiz$_4$**		**VizWiz$_2$**	
	Acc	Instby	Acc	Instby	Acc	Instby	Acc	Instby	Acc	Instby	Acc	Instby
Generation Evaluation												
BLIP-2$_F$	75.57	0.07	87.32	0.03	73.07	0.11	75.19	0.06	24.44	0.25	46.68	0.00
InstructBLIP$_F$	77.75	0.04	83.72	0.02	77.06	0.07	83.28	0.02	27.59	0.15	47.43	0.01
InstructBLIP$_V$	72.81	0.06	86.28	0.03	71.78	0.06	80.77	0.06	48.00	0.05	49.37	0.41
LLaVA-1.0-7B$_V$	35.92	0.45	14.38	0.69	19.48	0.60	25.90	0.20	14.91	0.93	45.20	0.43
LLaVA-1.0-7B$_{L_2}$	50.07	0.23	41.42	0.21	50.74	0.15	41.42	0.21	27.31	0.28	46.44	0.02
MiniGPT4	43.89	0.20	43.06	0.44	48.85	0.31	43.06	0.44	31.30	0.39	45.89	0.27
mPLUG-Owl	37.56	0.52	37.30	0.29	37.54	0.30	42.19	0.41	31.85	0.75	46.09	0.49
PandaGPT	34.43	0.61	26.32	0.06	27.63	0.05	29.07	0.02	31.11	0.31	29.75	0.84
ImageBindLLM	26.99	0.13	23.42	0.00	22.45	0.12	24.59	0.48	26.02	0.29	49.90	0.24
LA-V2	25.89	0.44	25.96	0.42	18.08	0.64	29.07	0.02	33.05	0.29	54.21	0.03
mmGPT	26.21	0.64	25.92	0.27	25.60	0.25	27.48	0.25	28.24	0.69	50.59	0.32
Shikra	41.54	0.11	50.72	0.11	47.99	0.10	42.62	0.11	21.48	0.21	47.72	0.13
Cheetor$_V$	55.26	0.27	59.12	0.11	46.51	0.25	44.86	0.28	33.98	0.23	49.90	0.16
Cheetor$_{L_2}$	37.80	0.23	70.82	0.15	43.64	0.15	36.39	0.21	31.11	0.24	46.88	0.06
BLIVA	30.71	0.22	37.52	0.21	36.68	0.20	35.57	0.25	32.78	0.19	48.71	0.20
LLaVA-1.5-7B$_V$	68.14	0.46	87.54	0.34	72.80	0.42	79.73	0.41	39.54	0.61	52.87	0.32
MiniGPT-v2	30.32	0.81	48.02	0.82	32.06	0.83	27.21	0.87	33.89	0.83	46.19	0.23
Qwen-VL-Chat	88.39	0.33	78.08	0.42	79.14	0.38	93.17	0.32	34.65	0.55	53.17	0.41
LLaVA-1.6-7B$_V$	70.68	0.41	84.16	0.37	74.44	0.41	84.10	0.41	33.61	0.71	58.71	0.45
Monkey	81.69	0.26	61.00	0.56	70.57	0.39	86.94	0.20	37.59	0.38	50.40	0.20
Deepseek-VL	80.12	0.90	82.34	0.91	76.65	0.88	79.23	0.88	38.70	0.79	50.59	0.55
ShareGPT4V-7B	64.40	0.33	76.92	0.42	69.68	0.32	82.35	0.24	38.80	0.54	51.29	0.42
shareGPT4V-13B	65.45	0.52	81.18	0.41	74.13	0.46	37.98	0.38	35.83	0.32	51.44	0.30
OmniLMM-12B	89.78	0.32	**88.02**	0.32	86.87	0.33	97.60	0.30	40.09	0.32	65.54	0.28
LLaVA-1.5-13B$_V$	69.63	0.42	80.12	0.38	74.88	0.43	86.45	0.37	38.80	0.34	52.13	0.22
LLaVA-1.6-13B$_V$	71.49	0.25	77.82	0.12	76.41	0.17	88.52	0.14	38.43	0.10	60.59	0.38
Qwen-VL-Max	93.77	-	76.30	-	**89.75**	-	98.63	-	37.96	-	**67.08**	-
Gemini-1.0-Pro-Vis	**95.23**	-	64.30	-	87.66	-	**99.73**	-	42.59	-	56.68	-
GPT-4V	93.99	-	58.65	-	86.93	-	95.36	-	**58.80**	-	66.34	-
Likelihood Evaluation												
BLIP-2$_F$	56.31	0.05	89.40	0.03	51.40	0.07	54.10	0.07	12.78	0.01	48.12	0.04
InstructBLIP$_F$	55.23	0.04	81.62	0.02	53.32	0.06	56.34	0.05	12.96	0.03	49.45	0.05
InstructBLIP$_V$	50.44	0.04	88.40	0.03	44.04	0.06	51.20	0.06	14.91	0.02	51.09	0.08
LLaVA-1.0-7B$_V$	48.78	0.04	92.56	0.01	51.16	0.05	47.98	0.05	14.35	0.00	54.21	0.06
LLaVA-1.0-7B$_{L_2}$	48.51	0.04	48.52	0.06	42.12	0.06	48.52	0.06	13.33	0.02	48.91	0.02
MiniGPT4	52.27	0.03	55.41	0.04	52.03	0.05	55.41	0.04	36.02	0.03	46.88	0.01
mPLUG-Owl	59.98	0.02	88.66	0.02	51.86	0.03	75.08	0.02	20.09	0.05	46.53	0.00
PandaGPT	48.78	0.06	76.86	0.06	43.89	0.08	24.21	0.11	27.03	0.11	48.02	0.00
ImageBindLLM	48.66	0.03	83.8	0.02	43.18	0.05	48.31	0.04	14.63	0.03	46.53	0.00
LA-V2	30.83	0.09	62.14	0.08	24.49	0.00	47.27	0.08	12.96	0.04	46.53	0.00
mmGPT	41.78	0.04	93.34	0.04	45.02	0.08	41.53	0.06	13.70	0.05	46.53	0.00
Shikra	50.86	0.01	89.70	0.02	47.99	0.10	53.77	0.04	28.70	0.03	57.48	0.04
Cheetor$_V$	49.36	0.05	87.30	0.03	47.88	0.09	43.06	0.11	22.22	0.10	50.30	0.05
Cheetor$_{L_2}$	45.35	0.07	96.88	0.02	38.13	0.09	39.07	0.10	18.89	0.06	46.58	0.02
BLIVA	59.36	0.04	94.78	0.01	58.27	0.04	67.10	0.04	19.35	0.03	48.02	0.03
LLaVA-1.5-7B$_V$	44.67	0.02	86.32	0.00	50.54	0.03	49.89	0.04	15.46	0.03	57.03	0.08
MiniGPT-v2	32.76	0.09	60.72	0.15	31.56	0.11	28.63	0.11	**43.33**	0.04	46.98	0.00
Qwen-VL-Chat	**76.50**	0.03	90.40	0.03	**64.70**	0.04	**90.00**	0.02	27.04	0.12	**63.32**	0.12
LLaVA-1.6-7B$_V$	49.10	0.03	86.28	0.00	53.02	0.03	59.23	0.03	14.72	0.00	57.62	0.05
Monkey	74.11	0.03	50.74	0.10	62.18	0.04	67.81	0.04	7.40	0.05	53.21	0.09
Deepseek-VL	55.31	0.06	68.10	0.02	50.78	0.08	55.85	0.10	17.69	0.02	51.73	0.03
ShareGPT4V-7B	47.16	0.02	96.18	0.01	54.16	0.04	54.21	0.04	12.96	0.02	58.17	0.08
shareGPT4V-13B	47.68	0.02	**97.84**	0.00	57.54	0.02	50.11	0.04	14.26	0.05	59.31	0.04
OmniLMM-12B	61.93	0.05	84.10	0.01	59.74	0.05	68.69	0.06	33.61	0.03	61.04	0.13
LLaVA-1.5-13B$_V$	47.65	0.02	86.68	0.00	53.57	0.03	53.44	0.05	16.76	0.01	58.96	0.04
LLaVA-1.6-13B$_V$	58.73	0.03	95.82	0.01	62.62	0.04	67.27	0.05	14.72	0.02	58.71	0.04
Qwen-VL-Max	-	-	-	-	-	-	-	-	-	-	-	-
Gemini-1.0-Pro-Vis	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4V	-	-	-	-	-	-	-	-	-	-	-	-

Table 15: Evaluation results on coarse-grained perception. "Acc" and "Instby" are short for accuracy and instability, respectively.

Model	**Scene Recognition**						**Avg.**	
	TDIUC$_{\textbf{sport}}$		**TDIUC$_{\textbf{scene}}$**		**MEDIC$_{\textbf{dts}}$**			
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
Generation Evaluation								
BLIP-2$_F$	93.75	0.12	88.66	0.04	60.29	0.25	69.44	0.10
InstructBLIP$_F$	93.79	0.12	89.27	0.05	60.76	0.15	71.18	0.07
InstructBLIP$_V$	90.62	0.20	69.78	0.09	52.00	0.13	69.06	0.12
LLaVA-1.0-7B$_V$	39.29	1.17	28.47	1.00	34.67	0.48	28.69	0.66
LLaVA-1.0-7B$_{L_2}$	74.13	0.51	67.29	0.16	36.00	0.28	48.31	0.23
MiniGPT4	65.41	0.68	58.04	0.40	36.19	0.44	46.19	0.40
mPLUG-Owl	59.54	0.85	58.79	0.64	26.67	0.81	41.95	0.56
PandaGPT	22.39	1.35	38.04	0.86	14.95	0.66	28.19	0.53
ImageBindLLM	24.59	1.33	45.70	0.59	19.43	0.77	29.23	0.44
LA-V2	42.94	1.11	48.69	0.54	20.76	0.79	33.18	0.48
mmGPT	28.81	1.29	45.23	0.64	15.25	0.87	30.37	0.58
Shikra	78.26	0.43	60.56	0.53	34.00	0.27	47.21	0.22
Cheetor$_V$	76.33	0.47	59.63	0.35	42.38	0.49	52.00	0.29
Cheetor$_{L_2}$	53.58	0.92	68.6	0.15	29.71	0.29	46.50	0.27
BLIVA	65.87	0.68	56.73	0.41	30.95	0.41	41.72	0.31
LLaVA-1.5-7B$_V$	89.17	0.22	78.13	0.27	50.76	0.65	68.74	0.41
MiniGPT-v2	78.17	0.41	58.32	0.35	58.32	1.10	45.83	0.69
Qwen-VL-Chat	89.82	0.10	87.66	0.12	53.14	0.50	73.02	0.35
LLaVA-1.6-7B$_V$	88.81	0.22	83.27	0.19	49.9	0.68	69.74	0.43
Monkey	91.56	0.16	80.93	0.17	60.67	0.31	69.04	0.29
Deepseek-VL	91.10	0.19	85.23	0.08	33.24	1.11	68.58	0.70
ShareGPT4V-7B	90.46	0.20	87.29	0.09	55.24	0.47	68.49	0.34
shareGPT4V-13B	88.72	0.22	87.94	0.06	56.19	0.53	64.32	0.36
OmniLMM-12B	93.67	0.14	**89.44**	0.02	58.57	0.42	78.84	0.27
LLaVA-1.5-13B$_V$	86.33	0.25	84.77	0.13	56.57	0.52	69.96	0.34
LLaVA-1.6-13B$_V$	91.19	0.18	88.41	0.06	66.10	0.27	73.22	0.18
Qwen-VL-Max	**94.95**	-	89.20	-	70.67	-	**79.81**	-
Gemini-1.0-Pro-Vis	91.74	-	88.79	-	72.38	-	77.68	-
GPT-4V	93.58	-	85.05	-	**74.24**	-	79.22	-
Likelihood Evaluation								
BLIP-2$_F$	96.32	0.05	89.90	0.02	48.00	0.05	60.70	0.04
InstructBLIP$_F$	96.9	0.05	91.21	0.01	46.10	0.09	60.35	0.04
InstructBLIP$_V$	97.2	0.04	90.98	0.01	38.57	0.52	58.54	0.10
LLaVA-1.0-7B$_V$	92.67	0.11	89.53	0.08	57.62	0.20	60.98	0.07
LLaVA-1.0-7B$_{L_2}$	76.33	0.32	80.47	0.19	42.76	0.40	49.94	0.13
MiniGPT4	87.52	0.19	68.88	0.12	39.62	0.37	54.89	0.10
mPLUG-Owl	80.91	0.28	64.02	0.27	34.19	0.28	57.92	0.11
PandaGPT	41.28	0.88	55.98	0.35	14.76	0.67	42.31	0.26
ImageBindLLM	78.53	0.35	55.61	0.33	26.95	0.25	49.58	0.12
LA-V2	71.28	0.42	63.93	0.22	24.57	0.49	42.67	0.16
mmGPT	82.11	0.27	71.87	0.07	37.24	0.36	52.57	0.11
Shikra	92.11	0.12	85.14	0.06	42.19	0.39	60.88	0.09
Cheetor$_V$	88.81	0.18	77.29	0.08	38.48	0.36	56.08	0.12
Cheetor$_{L_2}$	78.90	0.31	72.24	0.06	38.10	0.29	52.68	0.11
BLIVA	96.33	0.05	**93.83**	0.10	47.14	0.19	64.91	0.06
LLaVA-1.5-7B$_V$	97.80	0.04	90.56	0.01	47.33	0.18	59.96	0.05
MiniGPT-v2	97.25	0.04	65.79	0.15	25.24	0.71	48.03	0.16
Qwen-VL-Chat	98.53	0.02	92.43	0.01	24.95	0.43	**69.76**	0.09
LLaVA-1.6-7B$_V$	97.80	0.04	89.35	0.01	45.52	0.22	61.40	0.05
Monkey	96.79	0.05	87.20	0.03	47.43	0.41	60.76	0.09
Deepseek-VL	97.06	0.04	86.73	0.02	26.10	0.35	56.59	0.08
ShareGPT4V-7B	97.80	0.03	89.72	0.02	48.95	0.30	62.15	0.06
shareGPT4V-13B	**98.81**	0.02	90.28	0.01	60.00	0.11	63.98	0.03
OmniLMM-12B	98.44	0.03	89.25	0.03	49.05	0.17	67.32	0.06
LLaVA-1.5-13B$_V$	98.26	0.03	91.78	0.01	47.14	0.14	61.58	0.04
LLaVA-1.6-13B$_V$	97.71	0.03	90.37	0.01	**62.00**	0.17	67.55	0.04
Qwen-VL-Max	-	-	-	-	-	-	-	-
Gemini-1.0-Pro-Vis	-	-	-	-	-	-	-	-
GPT-4V	-	-	-	-	-	-	-	-

Table 16: Supplement of Table 15

Model	Object Perception									
	TDIUC$_{\text{color}}$		**TDIUC$_{\text{utility}}$**		**TDIUC$_{\text{position}}$**		**TDIUC$_{\text{detection}}$**		**TDIUC$_{\text{counting}}$**	
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
					Generation Evaluation					
BLIP-2$_F$	73.90	0.45	92.40	0.09	91.55	0.19	**99.38**	0.02	71.00	0.54
InstructBLIP$_F$	78.22	0.39	95.56	0.03	90.26	0.22	98.77	0.03	73.05	0.50
InstructBLIP$_V$	69.19	0.58	89.01	0.09	81.29	0.42	97.26	0.07	61.76	0.77
LLaVA-1.0-7B$_V$	8.25	1.46	57.89	0.80	30.60	1.33	53.29	0.97	23.43	1.43
LLaVA-1.0-7B$_{L_2}$	57.02	0.83	85.85	0.20	68.70	0.69	91.78	0.20	43.14	1.10
MiniGPT4	46.48	1.03	72.16	0.51	56.21	0.95	83.77	0.39	42.38	1.12
mPLUG-Owl	28.20	1.32	55.67	0.83	40.86	1.23	60.41	0.89	28.10	1.32
PandaGPT	30.18	1.27	57.19	0.79	27.59	1.38	61.99	0.85	26.52	1.33
ImageBindLLM	25.17	1.36	42.69	1.04	31.12	1.37	39.93	1.26	28.10	1.33
LA-V2	26.48	1.33	40.58	1.08	29.66	1.40	43.90	1.19	25.10	1.35
mmGPT	27.10	1.32	38.71	1.11	26.29	1.43	38.63	1.27	25.33	1.35
Shikra	39.43	1.15	61.64	0.69	50.43	1.00	61.23	0.85	27.67	0.43
Cheetor$_V$	47.10	1.01	78.01	0.34	49.91	1.05	85.34	0.34	35.71	1.22
Cheetor$_{L_2}$	54.57	0.84	81.52	0.29	57.76	0.90	84.72	0.37	31.29	1.26
BLIVA	44.28	1.06	61.75	0.68	45.09	1.14	63.49	0.82	41.95	1.13
LLaVA-1.5-7B$_V$	78.07	0.42	89.24	0.18	90.69	0.21	96.44	0.08	72.76	0.51
MiniGPT-v2	46.95	1.00	83.16	0.28	40.52	1.22	69.18	0.70	27.86	1.27
Qwen-VL-Chat	87.10	0.12	94.85	0.07	89.74	0.14	99.04	0.01	68.33	0.30
LLaVA-1.6-7B$_V$	69.09	0.54	91.58	0.10	86.72	0.29	96.92	0.07	72.90	0.52
Monkey	88.25	0.23	85.38	0.18	85.86	0.32	96.92	0.07	66.48	0.68
Deepseek-VL	85.53	0.28	93.10	0.12	89.22	0.25	95.00	0.12	72.76	0.48
ShareGPT4V-7B	87.00	0.26	92.05	0.13	85.86	0.31	96.64	0.08	72.29	0.52
shareGPT4V-13B	86.74	0.27	93.45	0.10	89.74	0.23	96.92	0.07	75.90	0.47
OmniLMM-12B	91.70	0.15	95.56	0.05	93.01	0.15	97.88	0.05	80.04	0.34
LLaVA-1.5-13B$_V$	80.94	0.36	91.70	0.13	66.72	0.66	93.63	0.15	74.81	0.44
LLaVA-1.6-13B$_V$	80.00	0.38	94.97	0.06	87.16	0.29	96.23	0.09	76.48	0.43
Qwen-VL-Max	**91.88**	-	**97.66**	-	**96.97**	-	98.63	-	**80.53**	-
Gemini-1.0-Pro-Vis	84.33	-	91.81	-	92.67	-	96.58	-	75.48	-
GPT-4V	86.42	-	94.15	-	94.40	-	94.18	-	72.62	-
					Likelihood Evaluation					
BLIP-2$_F$	79.83	0.34	91.35	0.10	94.66	0.12	98.84	0.03	82.43	0.31
InstructBLIP$_F$	90.46	0.16	93.80	0.07	94.91	0.11	99.45	0.01	82.95	0.29
InstructBLIP$_V$	91.64	0.15	91.35	0.14	95.78	0.10	99.11	0.02	79.86	0.33
LLaVA-1.0-7B$_V$	70.44	0.48	83.27	0.12	77.59	0.46	97.05	0.06	62.23	0.59
LLaVA-1.0-7B$_{L_2}$	55.72	0.72	82.81	0.21	71.90	0.59	91.10	0.20	74.71	0.44
MiniGPT4	69.92	0.49	86.90	0.13	76.12	0.49	96.37	0.09	73.19	0.46
mPLUG-Owl	57.75	0.68	90.41	0.10	75.69	0.52	93.22	0.15	71.00	0.49
PandaGPT	45.07	0.87	67.95	0.45	47.59	1.02	76.16	0.51	45.19	0.97
ImageBindLLM	36.14	0.92	82.22	0.25	63.53	0.73	81.16	0.39	23.95	1.09
LA-V2	55.72	0.72	88.70	0.15	69.57	0.64	83.84	0.35	54.33	0.7
mmGPT	44.96	0.85	86.32	0.14	74.74	0.54	95.82	0.10	63.81	0.62
Shikra	58.07	0.67	86.67	0.16	73.10	0.55	88.22	0.25	74.24	0.43
Cheetor$_V$	72.74	0.46	82.69	0.16	80.60	0.42	98.08	0.05	70.90	0.49
Cheetor$_{L_2}$	61.15	0.64	66.32	0.55	75.43	0.50	92.05	0.18	50.48	0.76
BLIVA	89.87	0.18	92.05	0.07	95.17	0.11	99.25	0.02	82.57	0.29
LLaVA-1.5-7B$_V$	92.58	0.13	92.51	0.07	95.52	0.10	**99.45**	0.01	81.71	0.32
MiniGPT-v2	87.94	0.22	92.51	0.13	59.14	0.82	82.87	0.37	80.19	0.35
Qwen-VL-Chat	92.95	0.04	96.84	0.04	92.24	0.05	98.08	0.01	83.00	0.10
LLaVA-1.6-7B$_V$	89.97	0.18	91.93	0.14	94.40	0.12	98.77	0.03	78.14	0.37
Monkey	93.00	0.12	**98.13**	0.03	92.85	0.16	98.29	0.04	83.10	0.30
Deepseek-VL	92.22	0.14	91.69	0.14	92.67	0.15	99.18	0.02	78.62	0.34
ShareGPT4V-7B	93.84	0.11	92.63	0.05	94.66	0.12	99.18	0.02	80.62	0.31
shareGPT4V-13B	93.68	0.11	94.74	0.06	**99.52**	0.01	82.86	0.01	82.86	0.31
OmniLMM-12B	**93.90**	0.12	93.68	0.10	94.22	0.13	98.97	0.03	**84.67**	0.27
LLaVA-1.5-13B$_V$	91.64	0.14	95.32	0.05	96.29	0.09	**99.45**	0.01	80.48	0.33
LLaVA-1.6-13B$_V$	91.96	0.15	92.87	0.10	94.83	0.12	99.32	0.02	80.10	0.35
Qwen-VL-Max	-	-	-	-	-	-	-	-	-	-
Gemini-1.0-Pro-Vis	-	-	-	-	-	-	-	-	-	-
GPT-4V	-	-	-	-	-	-	-	-	-	-

Table 17: Evaluation results on fine-grained perception.

Model	Object Perception								Object Grounding		Avg.	
	MSCOCO$_{\text{count}}$		MSCOCO$_{\text{mci}}$		MSCOCO$_{\text{goi}}$		MSCOCO$_{\text{mos}}$		RefCOCO$_{\text{res}}$			
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
Generation Evaluation												
BLIP-2$_F$	87.48	0.26	82.22	0.09	59.50	0.09	39.11	0.21	69.58	0.20	77.39	0.21
InstructBLIP$_F$	87.25	0.26	84.72	0.08	63.05	0.11	37.65	0.28	72.75	0.12	78.73	0.21
InstructBLIP$_V$	64.02	0.76	83.17	0.17	64.94	0.24	39.27	0.55	57.58	0.24	72.21	0.41
LLaVA-1.0-7B$_V$	18.60	1.48	44.88	0.91	29.56	1.14	34.66	0.94	42.41	0.38	33.46	1.16
LLaVA-1.0-7B$_{L_2}$	40.78	1.16	64.67	0.48	53.33	0.36	41.54	0.64	50.75	0.30	60.76	0.63
MiniGPT4	35.95	1.24	58.50	0.66	53.06	0.61	41.94	0.68	41.00	0.29	54.49	0.76
mPLUG-Owl	28.07	1.33	31.00	1.02	30.94	1.04	36.60	0.91	32.08	0.78	37.76	1.10
PandaGPT	22.92	1.40	25.11	1.03	25.16	1.01	41.54	0.70	28.08	0.47	35.36	1.08
ImageBindLLM	28.54	1.32	32.06	0.90	30.67	0.85	40.24	0.89	28.75	0.62	33.17	1.15
LA-V2	26.43	1.34	22.33	1.06	21.11	1.11	42.02	0.72	30.75	0.76	30.85	1.18
mmGPT	22.34	1.40	30.33	1.00	27.56	1.01	41.38	0.75	25.58	0.93	30.85	1.18
Shikra	29.04	1.31	70.50	0.44	54.61	0.57	28.91	0.57	51.75	0.22	47.05	0.78
Cheetor$_V$	29.63	1.29	49.67	0.78	45.11	0.72	38.78	0.70	43.75	0.52	51.03	0.83
Cheetor$_{L_2}$	30.57	1.29	52.16	0.66	43.16	0.66	40.32	0.67	37.42	0.46	52.90	0.77
BLIVA	37.93	1.20	39.94	0.82	32.22	0.89	40.00	0.67	27.58	0.35	45.18	0.93
LLaVA-1.5-7B$_V$	82.61	0.36	84.17	0.03	67.72	0.05	44.45	0.12	64.08	0.51	78.46	0.22
MiniGPT-v2	27.88	1.32	35.44	0.12	34.78	0.14	**90.36**	0.03	27.33	0.88	50.68	0.68
Qwen-VL-Chat	68.73	0.32	84.44	0.08	72.44	0.11	41.54	0.35	67.92	0.47	78.47	0.17
LLaVA-1.6-7B$_V$	79.92	0.41	84.61	0.13	66.44	0.11	44.94	0.30	68.50	0.45	77.01	0.27
Monkey	62.26	0.81	87.56	0.14	68.61	0.27	39.11	0.32	63.30	0.49	75.60	0.34
Deepseek-VL	90.53	0.19	87.01	0.12	80.33	0.15	39.35	0.30	70.66	0.45	81.43	0.22
ShareGPT4V-7B	81.33	0.40	84.72	0.15	71.22	0.19	37.33	0.47	78.75	0.22	78.72	0.28
ShareGPT4V-13B	84.05	0.33	88.89	0.10	73.89	0.12	39.92	0.39	71.50	0.45	81.06	0.23
OmniLMM-12B	94.23	0.12	92.78	0.04	78.72	0.07	39.60	0.35	74.42	0.41	84.84	0.15
LLaVA-1.5-13B$_V$	83.98	0.33	89.56	0.06	69.28	0.16	30.20	0.23	67.33	0.45	75.65	0.28
LLaVA-1.6-13B$_V$	84.25	0.31	86.50	0.09	70.06	0.08	41.54	0.24	76.58	0.23	79.69	0.22
Qwen-VL-Max	**94.74**	-	**94.71**	-	88.86	-	37.40	-	**90.00**	-	**86.82**	-
Gemini-1.0-Pro-Vis	89.47	-	86.94	-	**89.17**	-	54.66	-	81.67	-	84.57	-
GPT-4V	88.69	-	84.51	-	82.40	-	65.70	-	87.08	-	84.79	-
Likelihood Evaluation												
BLIP-2$_F$	77.31	0.41	81.77	0.02	60.39	0.02	39.27	0.02	38.58	0.15	78.43	0.15
InstructBLIP$_F$	73.14	0.48	82.94	0.04	60.44	0.05	42.27	0.06	36.08	0.09	80.04	0.14
InstructBLIP$_V$	89.98	0.17	84.72	0.02	64.94	0.02	43.89	0.04	37.17	0.11	82.36	0.11
LLaVA-1.0-7B$_V$	91.07	0.16	76.22	0.12	62.50	0.15	39.92	0.15	43.00	0.15	73.37	0.25
LLaVA-1.0-7B$_{L_2}$	89.01	0.20	64.44	0.17	47.44	0.12	40.24	0.13	38.33	0.09	68.60	0.31
MiniGPT4	87.56	0.22	77.11	0.06	58.22	0.06	41.62	0.07	38.50	0.10	74.11	0.23
mPLUG-Owl	89.04	0.20	56.61	0.13	49.56	0.14	38.38	0.10	39.33	0.10	69.07	0.28
PandaGPT	66.04	0.60	32.22	0.29	26.94	0.31	42.11	0.28	24.50	0.13	49.92	0.59
ImageBindLLM	87.06	0.22	48.61	0.09	41.78	0.13	43.72	0.09	35.92	0.10	56.46	0.43
LA-V2	87.45	0.22	49.83	0.08	46.83	0.16	41.21	0.09	36.67	0.11	64.16	0.35
mmGPT	69.12	0.56	61.56	0.11	52.19	0.37	43.08	0.10	32.33	0.08	65.73	0.38
Shikra	90.21	0.17	51.56	0.10	54.67	0.10	41.70	0.08	49.92	0.11	68.72	0.28
Cheetor$_V$	77.66	0.40	74.67	0.08	55.67	0.12	43.07	0.14	33.42	0.16	72.90	0.26
Cheetor$_{L_2}$	80.16	0.36	67.88	0.09	51.56	0.11	39.92	0.12	32.00	0.13	64.99	0.37
BLIVA	94.04	0.10	84.00	0.03	66.72	0.04	42.35	0.06	35.75	0.13	82.89	0.10
LLaVA-1.5-7B$_V$	90.64	0.17	84.44	0.00	66.94	0.00	**50.20**	0.00	42.83	0.11	83.78	0.09
MiniGPT-v2	70.92	0.53	38.61	0.00	37.50	0.00	46.96	0.00	27.50	0.17	66.29	0.27
Qwen-VL-Chat	94.19	0.02	81.67	0.00	70.28	0.00	42.51	0.00	38.33	0.13	83.53	0.03
LLaVA-1.6-7B$_V$	90.06	0.18	82.78	0.00	67.22	0.00	43.72	0.00	40.75	0.14	81.89	0.11
Monkey	93.80	0.12	70.83	0.00	56.11	0.00	42.91	0.00	51.25	0.13	81.00	0.08
Deepseek-VL	93.22	0.12	77.61	0.00	75.22	0.00	41.30	0.00	48.50	0.18	82.41	0.10
ShareGPT4V-7B	92.98	0.13	85.00	0.00	74.72	0.00	46.56	0.00	**52.17**	0.11	84.47	0.08
ShareGPT4V-13B	91.15	0.16	87.22	0.00	76.39	0.00	43.32	0.00	28.17	0.09	83.53	0.07
OmniLMM-12B	**97.70**	0.05	**87.78**	0.00	**82.50**	0.00	43.32	0.00	47.17	0.10	**86.30**	0.08
LLaVA-1.5-13B$_V$	91.38	0.16	86.39	0.00	64.72	0.00	48.58	0.00	44.42	0.12	83.81	0.09
LLaVA-1.6-13B$_V$	89.63	0.19	85.00	0.00	69.44	0.00	46.96	0.00	49.42	0.12	83.35	0.10
Qwen-VL-Max	-	-	-	-	-	-	-	-	-	-	-	
Gemini-1.0-Pro-Vis	-	-	-	-	-	-	-	-	-	-	-	
GPT-4V	-	-	-	-	-	-	-	-	-	-	-	

Table 18: Supplement of Table 17

Model	VQA						KVQA					
	GQA		VQA v2		Whoops		OK-VQA		ScienceQA		VizWiz	
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
Generation Evaluation												
BLIP-2$_F$	62.66	0.18	69.86	0.19	81.96	0.10	68.97	0.23	63.78	0.27	82.13	0.12
InstructBLIP$_F$	66.11	0.19	74.31	0.16	80.04	0.08	70.87	0.15	62.19	0.26	74.90	0.18
InstructBLIP$_V$	59.09	0.31	62.59	0.32	71.96	0.27	63.73	0.36	58.01	0.42	49.37	0.41
LLaVA-1.0-7B$_V$	37.26	0.80	46.01	0.55	50.48	0.69	41.59	0.78	46.57	0.63	31.37	0.80
LLaVA-1.0-7B$_{L_2}$	52.36	0.43	51.65	0.40	57.14	0.43	54.09	0.57	57.91	0.49	40.32	0.54
MiniGPT4	44.57	0.66	46.49	0.64	47.08	0.78	38.65	0.88	43.78	0.71	35.22	0.83
mPLUG-Owl	34.56	0.89	35.48	0.88	37.44	0.93	33.45	0.97	41.39	0.80	30.63	0.96
PandaGPT	38.07	0.67	37.28	0.68	24.64	0.85	29.80	0.90	44.48	0.69	24.87	0.89
ImageBindLLM	38.63	0.75	38.56	0.77	27.86	0.95	31.67	0.96	41.49	0.73	26.22	0.97
LA-V2	40.21	0.68	39.27	0.67	34.17	0.94	29.52	1.00	41.59	0.76	28.63	0.93
mmGPT	35.12	0.85	34.47	0.85	27.20	1.01	27.34	1.00	40.10	0.78	23.67	0.95
Shikra	41.69	0.73	44.93	0.67	50.48	0.73	41.15	0.86	38.61	0.72	41.81	0.81
Cheetor$_V$	46.17	0.58	48.17	0.56	55.71	0.64	43.49	0.78	47.06	0.63	37.59	0.78
Cheetor$_{L_2}$	48.39	0.42	45.62	0.43	42.26	0.51	44.64	0.61	56.12	0.50	32.76	0.65
BLIVA	43.40	0.58	50.06	0.01	46.31	0.76	36.75	0.76	42.09	0.65	35.64	0.80
LLaVA-1.5-7B$_V$	63.33	0.33	70.46	0.29	78.51	0.17	72.94	0.24	61.69	0.35	56.47	0.30
MiniGPT-v2	43.48	0.55	40.73	0.52	31.61	0.68	37.66	0.68	53.93	0.53	21.25	0.76
Qwen-VL-Chat	67.33	0.31	73.37	0.30	78.81	0.12	69.56	0.16	61.89	0.38	67.66	0.25
LLaVA-1.6-7B$_V$	63.77	0.28	71.79	0.25	77.14	0.17	71.75	0.23	63.58	0.34	60.88	0.28
Monkey	67.53	0.32	73.93	0.26	80.06	0.21	68.57	0.38	60.30	0.39	72.67	0.33
Deepseek-VL	70.61	0.19	80.33	0.14	81.55	0.16	73.17	0.19	75.42	0.24	71.55	0.27
ShareGPT4V-7B	64.93	0.27	74.12	0.22	77.02	0.22	69.29	0.30	59.00	0.34	56.29	0.33
ShareGPT4V-13B	70.36	0.19	79.32	0.16	81.25	0.17	74.00	0.25	66.17	0.30	70.86	0.24
OmniLMM-12B	74.34	0.11	84.05	0.09	**85.36**	0.11	79.25	0.15	76.32	0.20	80.65	0.16
LLaVA-1.5-13B$_V$	47.21	0.54	74.10	0.18	78.27	0.15	72.82	0.19	68.66	0.27	72.30	0.23
LLaVA-1.6-13B$_V$	64.39	0.22	75.28	0.14	79.82	0.13	75.12	0.18	70.05	0.29	77.08	0.19
Qwen-VL-Max	**77.55**	-	**85.79**	-	81.96	-	**82.27**	-	71.14	-	**93.50**	-
Gemini-1.0-Pro-Vis	71.12	-	81.34	-	84.82	-	76.39	-	**78.61**	-	84.45	-
GPT-4V	62.53	-	73.09	-	75.30	-	71.03	-	70.15	-	84.45	-
Likelihood Evaluation												
BLIP-2$_F$	62.70	0.06	69.37	0.06	70.95	0.04	66.83	0.07	53.93	0.03	76.71	0.08
InstructBLIP$_F$	66.65	0.06	79.67	0.05	68.99	0.04	76.35	0.03	56.32	0.03	62.92	0.04
InstructBLIP$_V$	67.76	0.04	**82.92**	0.02	72.32	0.02	**82.82**	0.02	57.91	0.02	57.17	0.03
LLaVA-1.0-7B$_V$	54.14	0.12	58.94	0.09	65.36	0.06	62.18	0.11	54.03	0.08	40.28	0.07
LLaVA-1.0-7B$_{L_2}$	52.68	0.09	51.81	0.10	60.06	0.06	48.77	0.15	55.22	0.09	47.05	0.09
MiniGPT4	56.10	0.06	56.04	0.06	63.15	0.05	55.08	0.10	51.74	0.04	49.00	0.05
mPLUG-Owl	50.95	0.06	50.53	0.07	60.89	0.05	49.80	0.10	50.25	0.04	44.32	0.09
PandaGPT	41.35	0.20	37.86	0.25	28.69	0.20	36.11	0.24	43.88	0.13	19.12	0.17
ImageBindLLM	46.86	0.03	47.06	0.03	48.93	0.04	45.52	0.05	54.03	0.03	32.71	0.05
LA-V2	50.29	0.06	47.36	0.05	60.71	0.06	44.60	0.10	50.05	0.03	41.21	0.08
mmGPT	52.86	0.06	48.54	0.06	49.64	0.06	56.43	0.10	49.25	0.03	38.93	0.06
Shikra	57.07	0.09	64.38	0.07	65.83	0.06	59.72	0.08	51.54	0.06	42.00	0.04
Cheetor$_V$	58.71	0.09	59.70	0.09	62.86	0.08	58.81	0.11	48.26	0.08	44.73	0.13
Cheetor$_{L_2}$	53.03	0.08	50.13	0.10	56.55	0.09	50.63	0.13	55.42	0.06	35.96	0.11
BLIVA	67.37	0.05	81.36	0.03	69.88	0.03	78.53	0.03	60.70	0.02	51.32	0.04
LLaVA-1.5-7B$_V$	69.39	0.10	69.56	0.09	66.85	0.06	67.14	0.12	58.11	0.07	54.99	0.07
MiniGPT-v2	42.50	0.27	42.71	0.25	36.19	0.19	40.91	0.30	48.16	0.14	28.17	0.18
Qwen-VL-Chat	72.16	0.14	78.33	0.13	**75.60**	0.10	73.65	0.22	65.17	0.13	85.99	0.08
LLaVA-1.6-7B$_V$	66.05	0.16	65.50	0.13	67.14	0.08	65.28	0.14	56.32	0.07	53.69	0.08
Monkey	74.05	0.04	82.80	0.03	72.98	0.03	80.60	0.03	53.13	0.05	**88.54**	0.02
Deepseek-VL	72.20	0.11	76.52	0.10	73.51	0.09	68.93	0.12	**74.63**	0.07	84.32	0.06
ShareGPT4V-7B	73.37	0.09	73.04	0.07	69.35	0.05	69.88	0.11	57.61	0.06	59.44	0.05
ShareGPT4V-13B	**74.80**	0.05	77.52	0.05	73.21	0.05	77.22	0.07	60.50	0.05	65.10	0.06
OmniLMM-12B	64.61	0.13	71.23	0.11	69.35	0.08	65.24	0.16	68.46	0.08	84.78	0.09
LLaVA-1.5-13B$_V$	73.33	0.08	74.16	0.06	69.58	0.05	74.72	0.11	60.00	0.05	61.86	0.07
LLaVA-1.6-13B$_V$	68.39	0.05	70.52	0.10	70.60	0.06	70.79	0.10	62.69	0.04	56.01	0.07
Qwen-VL-Max	-	-	-	-	-	-	-	-	-	-	-	-
Gemini-1.0-Pro-Vis	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4V	-	-	-	-	-	-	-	-	-	-	-	-

Table 19: Evaluation results on visually grounded reasoning.

Model	KVQA												Avg.	
	ViQuAE		K-ViQuAE		A-OKVQA		A-OKVQRA		A-OKVQAR		ImageNetVC			
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
Generation Evaluation														
BLIP-2_F	44.32	0.37	87.26	0.10	71.40	0.18	85.61	0.07	80.00	0.17	81.72	0.14	73.82	0.18
InstructBLIP$_F$	43.68	0.36	87.10	0.09	77.02	0.17	89.82	0.05	81.23	0.16	79.07	0.14	72.47	0.17
InstructBLIP$_V$	49.76	0.41	80.32	0.26	71.58	0.28	79.47	0.24	47.72	0.61	62.56	0.30	66.15	0.31
LLaVA-1.0-7B$_V$	39.52	0.77	50.00	0.73	42.56	0.78	62.11	0.62	31.58	0.83	48.94	0.49	47.24	0.67
LLaVA-1.0-7B$_{L_2}$	47.20	0.50	85.00	0.17	60.18	0.47	81.93	0.24	61.93	0.40	66.58	0.30	63.98	0.35
MiniGPT4	32.48	0.91	65.81	0.56	39.65	0.85	59.30	0.67	40.18	0.82	53.81	0.57	49.80	0.71
mPLUG-Owl	31.04	0.99	51.29	0.81	35.26	1.01	41.58	0.88	35.61	0.90	42.65	0.79	40.61	0.88
PandaGPT	39.84	0.74	74.68	0.37	29.82	0.94	63.16	0.62	39.82	0.76	56.81	0.55	48.99	0.63
ImageBindLLM	29.28	0.97	46.13	0.86	29.65	0.96	54.56	0.76	32.28	1.01	44.32	0.70	36.90	0.87
LA-V2	27.20	1.00	40.32	0.90	31.75	0.99	47.02	0.86	31.93	0.90	43.78	0.56	36.37	0.85
mmGPT	31.04	0.99	45.81	0.90	24.39	1.00	38.42	0.93	25.26	0.95	43.14	0.73	36.80	0.91
Shikra	29.92	0.95	38.71	0.86	41.93	0.83	40.53	0.86	37.19	0.78	47.13	0.65	41.56	0.80
Cheetor$_V$	40.48	0.80	70.00	0.48	41.05	0.76	63.33	0.59	48.42	0.68	56.81	0.51	55.75	0.61
Cheetor$_{L_2}$	44.16	0.57	82.26	0.22	48.42	0.54	82.28	0.22	57.37	0.50	68.50	0.26	59.30	0.39
BLIVA	33.92	0.76	45.16	0.85	44.21	0.71	54.74	0.74	31.05	0.77	45.70	0.56	42.77	0.73
LLaVA-1.5-7B$_V$	57.92	0.28	88.71	0.10	80.35	0.16	91.23	0.10	78.78	0.23	69.09	0.28	73.56	0.21
MiniGPT-v2	33.60	0.73	81.13	0.27	40.18	0.75	73.51	0.36	34.74	0.71	62.16	0.33	52.13	0.50
Qwen-VL-Chat	59.20	0.22	81.94	0.10	75.96	0.27	89.12	0.12	70.88	0.35	74.64	0.21	73.65	0.16
LLaVA-1.6-7B$_V$	61.28	0.24	90.00	0.11	73.68	0.17	91.05	0.10	74.21	0.27	72.24	0.24	75.17	0.19
Monkey	54.40	0.43	81.94	0.23	75.79	0.23	89.30	0.12	56.84	0.55	76.56	0.21	73.24	0.27
Deepseek-VL	53.76	0.33	88.38	0.11	78.60	0.18	91.05	0.11	72.28	0.32	77.54	0.13	75.31	0.18
ShareGPT4V-7B	56.16	0.36	88.71	0.11	73.51	0.21	89.12	0.13	71.05	0.27	75.43	0.19	74.33	0.22
ShareGPT4V-13B	58.40	0.29	92.42	0.07	77.37	0.19	93.51	0.05	80.35	0.17	77.74	0.15	77.45	0.17
OmniLMM-12B	62.08	0.23	92.90	0.06	82.81	0.11	93.86	0.06	85.26	0.18	83.93	0.08	81.07	0.12
LLaVA-1.5-13B$_V$	66.70	0.26	93.87	0.05	76.32	0.18	92.98	0.05	81.75	0.17	68.99	0.30	76.96	0.19
LLaVA-1.6-13B$_V$	65.44	0.25	93.06	0.07	80.70	0.00	92.98	0.04	77.02	0.21	77.74	0.13	79.02	0.15
Qwen-VL-Max	**83.47**	-	95.87	-	**85.96**	-	**95.61**	-	87.72	-	84.58	-	**86.47**	-
Gemini-1.0-Pro-Vis	80.80	-	92.74	-	84.21	-	**95.61**	-	82.46	-	**87.22**	-	86.40	-
GPT-4V	73.95	-	**96.67**	-	74.56	-	91.22	-	**88.60**	-	85.26	-	82.80	-
Likelihood Evaluation														
BLIP-2_F	38.72	0.10	87.26	0.01	64.74	0.08	81.58	0.03	**84.04**	0.01	80.34	0.07	69.32	0.06
InstructBLIP$_F$	33.12	0.05	**88.71**	0.02	70.88	0.06	79.82	0.03	83.86	0.01	84.47	0.05	70.98	0.04
InstructBLIP$_V$	46.88	0.05	82.26	0.01	78.25	0.01	86.67	0.04	81.40	0.03	**85.70**	0.02	71.79	0.03
LLaVA-1.0-7B$_V$	39.20	0.10	75.00	0.07	59.30	0.09	73.51	0.12	61.40	0.05	63.34	0.07	60.73	0.08
LLaVA-1.0-7B$_{L_2}$	35.36	0.17	80.48	0.06	48.25	0.14	64.74	0.09	68.95	0.01	67.17	0.10	56.71	0.10
MiniGPT4	27.68	0.08	73.23	0.03	57.72	0.07	70.35	0.06	64.21	0.03	62.80	0.05	56.72	0.05
mPLUG-Owl	30.24	0.09	78.87	0.08	42.98	0.10	63.51	0.09	67.89	0.02	61.03	0.06	57.76	0.07
PandaGPT	24.64	0.26	77.26	0.07	31.75	0.25	59.30	0.14	61.93	0.05	57.44	0.17	47.01	0.18
ImageBindLLM	32.80	0.05	67.42	0.08	44.39	0.04	58.42	0.10	66.32	0.04	58.92	0.02	52.02	0.05
LA-V2	39.20	0.05	70.00	0.08	43.51	0.07	70.88	0.07	66.49	0.03	64.28	0.05	58.55	0.06
mmGPT	33.44	0.10	82.58	0.07	49.47	0.09	69.47	0.10	77.54	0.01	66.19	0.05	57.96	0.07
Shikra	35.20	0.10	65.65	0.04	57.72	0.07	64.74	0.11	75.26	0.00	62.56	0.09	57.31	0.07
Cheetor$_V$	34.40	0.10	70.81	0.07	59.12	0.11	70.88	0.09	70.18	0.02	66.29	0.06	58.73	0.09
Cheetor$_{L_2}$	37.12	0.13	82.26	0.06	53.33	0.10	73.51	0.08	74.91	0.02	66.78	0.11	60.68	0.10
BLIVA	51.84	0.06	83.87	0.02	**80.53**	0.03	**87.72**	0.01	79.12	0.02	82.95	0.04	72.14	0.04
LLaVA-1.5-7B$_V$	34.72	0.10	78.71	0.05	61.40	0.14	78.95	0.08	70.18	0.02	73.81	0.10	63.52	0.08
MiniGPT-v2	31.20	0.21	74.19	0.11	40.53	0.23	53.33	0.23	74.04	0.05	55.77	0.26	49.34	0.19
Qwen-VL-Chat	53.12	0.21	82.58	0.08	71.93	0.23	87.02	0.03	67.02	0.05	80.98	0.14	73.07	0.13
LLaVA-1.6-7B$_V$	41.44	0.11	75.97	0.05	60.88	0.15	78.07	0.08	69.47	0.03	70.52	0.15	63.77	0.10
Monkey	**58.24**	0.09	83.71	0.03	76.14	0.03	86.49	0.03	65.08	0.07	82.00	0.04	**74.23**	0.05
Deepseek-VL	40.32	0.15	78.71	0.05	66.49	0.12	85.26	0.08	71.75	0.05	75.82	0.10	67.09	0.10
ShareGPT4V-7B	37.28	0.12	78.55	0.09	66.49	0.09	80.70	0.07	70.18	0.02	75.72	0.06	65.23	0.08
ShareGPT4V-13B	50.72	0.09	80.65	0.08	75.26	0.09	81.75	0.05	73.16	0.02	80.29	0.05	71.22	0.07
OmniLMM-12B	49.60	0.15	81.77	0.09	61.75	0.16	85.96	0.07	75.61	0.05	82.51	0.08	70.81	0.10
LLaVA-1.5-13B$_V$	48.16	0.10	81.13	0.06	73.16	0.08	82.28	0.06	71.40	0.03	79.71	0.07	69.65	0.07
LLaVA-1.6-13B$_V$	48.64	0.11	79.84	0.08	61.23	0.13	79.12	0.09	70.53	0.05	74.35	0.09	68.36	0.08
Qwen-VL-Max	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Gemini-1.0-Pro-Vis	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4V	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 20: Supplement of Table 19.

Model	OCR						Grounded OCR		
	CUTE80	IC15	IIIT5K	COCO-Text	WordArt	TextOCR	gIC15	gCOCO-Text	gTextOCR
				Generation Evaluation					
BLIP-2$_F$	80.07	64.75	72.27	50.54	72.32	68.40	61.43	16.27	32.60
InstructBLIP$_F$	84.17	75.36	83.40	56.66	75.76	74.27	55.24	17.33	32.60
InstructBLIP$_V$	81.60	73.15	77.20	53.00	74.17	70.60	53.33	18.93	32.13
LLaVA-1.0-7B$_V$	26.11	23.65	25.73	13.00	34.44	36.00	40.24	17.33	30.00
LLaVA-1.0-7B$_{L_2}$	32.15	25.75	27.20	15.83	40.13	39.87	43.57	17.93	32.07
MiniGPT4	71.39	58.90	71.67	40.36	72.45	60.27	44.05	11.00	28.60
mPLUG-Owl	73.68	60.77	74.67	46.39	73.25	64.27	64.52	20.27	34.67
PandaGPT	1.60	1.88	3.13	0.14	5.03	26.87	0.24	0.40	22.73
ImageBindLLM	11.94	8.95	8.60	2.41	11.66	29.73	6.19	3.33	23.80
LA-V2	36.53	31.82	35.33	17.82	40.13	42.13	47.86	18.07	33.20
mmGPT	26.94	23.09	18.80	13.43	31.13	36.20	26.43	12.73	28.80
Shikra	2.57	4.75	5.07	4.33	9.54	31.13	29.76	5.93	27.53
Cheetor$_V$	52.50	39.01	53.87	29.73	56.16	52.27	40.00	11.20	28.20
Cheetor$_{L_2}$	42.78	31.38	39.20	20.36	34.83	45.40	16.67	6.67	25.13
BLIVA	77.29	68.40	72.47	51.49	71.26	66.93	64.76	21.67	37.27
LLaVA-1.5-7B$_V$	37.86	27.41	28.73	17.23	38.68	41.20	47.75	23.53	36.20
MiniGPT-v2	8.57	2.43	3.80	0.71	6.36	27.33	2.24	0.60	23.27
Qwen-VL-Chat	70.00	33.81	60.60	31.41	54.97	53.40	43.09	20.47	30.33
LLaVA-1.6-7B$_V$	46.43	29.72	30.20	18.73	34.30	40.67	58.56	25.93	35.47
Monkey	71.43	35.91	61.33	29.73	58.28	52.67	52.77	24.67	38.67
Deepseek-VL	82.14	44.20	75.73	38.16	71.52	62.33	70.65	33.67	43.87
ShareGPT4V-7B	46.43	27.29	31.20	16.00	36.56	42.27	61.60	30.47	35.73
ShareGPT4V-13B	41.43	32.49	33.67	18.62	34.57	42.67	65.42	30.00	38.27
OmniLMM-12B	85.71	76.80	91.73	61.58	82.78	76.67	76.75	44.80	47.33
LLaVA-1.5-13B$_V$	39.29	32.04	34.67	20.42	43.71	40.67	54.42	25.67	37.00
LLaVA-1.6-13B$_V$	53.57	31.49	30.67	20.12	39.74	40.33	56.89	25.33	40.33
Qwen-VL-Max	**96.43**	**81.77**	**98.00**	**68.96**	**86.09**	**81.61**	**88.38**	40.94	57.53
Gemini-1.0-Pro-Vis	92.86	51.93	82.33	30.83	78.81	64.67	85.98	**55.00**	**63.67**
GPT-4V	**96.43**	59.22	90.60	44.92	76.92	69.97	82.28	43.39	50.34

Table 21: Evaluation results on scene text perception.

Model	KIE			OCR-based VQA			Avg.
	FUNSD	POIE	SROIE	TextVQA	DocVQA	OCR-VQA	
Generation Evaluation							
BLIP-2$_F$	1.30	0.76	1.72	21.47	5.39	21.62	40.00
InstructBLIP$_F$	0.87	0.44	2.07	26.76	4.78	28.07	42.45
InstructBLIP$_V$	3.48	0.82	1.72	30.22	6.21	34.37	41.30
LLaVA-1.0-7B$_V$	0.00	0.44	1.72	19.02	3.13	5.74	17.27
LLaVA-1.0-7B$_{L_2}$	0.00	1.21	1.72	26.31	6.52	12.13	19.81
MiniGPT4	0.29	0.85	2.07	17.29	3.95	12.49	36.18
mPLUG-Owl	4.06	1.58	1.72	30.71	8.40	37.87	40.29
PandaGPT	0.00	0.09	1.72	0.80	2.22	0.00	1.77
ImageBindLLM	0.00	0.06	1.72	10.09	3.62	0.91	5.86
LA-V2	0.72	3.16	1.72	30.40	8.06	16.40	22.81
mmGPT	0.00	1.33	1.72	21.07	4.78	4.47	14.91
Shikra	0.00	0.82	1.72	1.56	0.19	0.25	6.05
Cheetor$_V$	0.14	0.79	1.72	13.16	3.62	7.26	27.53
Cheetor$_{L_2}$	0.00	0.57	1.72	11.02	4.11	3.05	18.98
BLIVA	2.61	3.04	3.45	29.69	6.18	34.97	41.05
LLaVA-1.5-7B$_V$	1.43	2.72	1.01	35.91	6.59	25.99	21.35
MiniGPT-v2	1.16	0.85	0.00	4.18	1.77	0.00	3.02
Qwen-VL-Chat	20.65	19.15	36.23	52.76	43.20	55.38	42.41
LLaVA-1.6-7B$_V$	1.73	5.57	4.35	37.33	8.25	24.01	24.35
Monkey	25.34	27.37	40.57	62.44	47.27	67.51	46.70
Deepseek-VL	7.35	34.49	18.12	61.02	30.96	56.14	48.35
ShareGPT4V-7B	3.71	17.44	17.54	48.67	10.96	27.01	28.08
ShareGPT4V-13B	4.01	18.92	12.31	45.82	14.09	29.39	28.55
OmniLMM-12B	4.97	19.33	7.97	58.89	25.05	51.68	52.34
LLaVA-1.5-13B$_V$	1.19	3.32	1.45	38.00	10.36	30.20	24.49
LLaVA-1.6-13B$_V$	2.04	7.44	4.35	35.56	9.04	27.41	26.14
Qwen-VL-Max	**55.10**	42.25	**65.94**	**74.67**	**85.31**	**72.12**	**77.70**
Gemini-1.0-Pro-Vis	39.97	47.31	48.55	67.11	65.35	54.82	65.90
GPT-4V	44.90	**64.24**	54.35	67.71	79.85	55.84	72.09

Table 22: Supplement of Table 21.

Model	Space-based Perception		Spatial Relation Judgment				Avg.	
	CLEVR		VSR		MP3D-Spatial			
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
Generation Evaluation								
BLIP-2$_F$	42.67	0.28	46.95	0.21	39.87	0.32	43.16	0.27
InstructBLIP$_F$	44.84	0.39	52.37	0.25	41.01	0.37	46.07	0.34
InstructBLIP$_V$	46.32	0.51	52.37	0.49	34.59	0.50	44.43	0.50
LLaVA-1.0-7B$_V$	19.01	1.24	40.00	0.88	27.19	1.13	28.73	1.08
LLaVA-1.0-7B$_{L_2}$	36.52	0.61	52.54	0.21	34.67	0.64	41.24	0.49
MiniGPT4	33.74	0.84	36.44	0.81	33.62	0.84	34.60	0.83
mPLUG-Owl	27.48	1.01	28.81	0.97	24.23	1.04	26.84	1.01
PandaGPT	29.65	0.90	35.76	0.86	34.50	0.80	33.30	0.85
ImageBindLLM	31.45	0.96	40.00	0.94	35.22	0.83	35.56	0.91
LA-V2	21.39	1.05	23.05	1.04	27.06	1.01	23.83	1.03
mmGPT	22.26	1.13	28.98	1.01	29.30	0.98	26.85	1.04
Shikra	23.82	0.77	46.27	0.60	29.77	0.84	33.29	0.74
Cheetor$_V$	24.72	1.03	35.76	0.77	31.21	0.88	30.56	0.89
Cheetor$_{L_2}$	29.10	0.77	40.85	0.69	33.53	0.73	34.49	0.73
BLIVA	30.64	0.85	35.25	0.61	34.12	0.59	33.34	0.68
LLaVA-1.5-7B$_V$	24.23	0.19	56.27	0.12	46.38	0.38	42.29	0.23
MiniGPT-v2	10.06	0.49	54.07	0.46	27.86	0.52	30.66	0.49
Qwen-VL-Chat	43.68	0.30	53.73	0.21	36.36	0.00	44.59	0.17
LLaVA-1.6-7B$_V$	40.92	0.42	58.31	0.39	45.67	0.44	48.30	0.42
Monkey	45.10	0.56	56.27	0.45	34.38	0.62	45.25	0.54
Deepseek-VL	49.01	0.37	55.42	0.38	45.62	0.47	50.02	0.41
ShareGPT4V-7B	43.62	0.51	58.47	0.48	42.92	0.57	48.34	0.52
ShareGPT4V-13B	52.03	0.29	66.95	0.27	48.25	0.44	55.74	0.33
OmniLMM-12B	**72.52**	0.11	**72.88**	0.16	**52.52**	0.27	**65.97**	0.18
LLaVA-1.5-13B$_V$	42.38	0.39	67.29	0.27	48.71	0.38	52.79	0.34
LLaVA-1.6-13B$_V$	48.49	0.29	66.61	0.26	45.58	0.37	53.56	0.31
Qwen-VL-Max	55.22	-	70.34	-	49.89	-	58.48	-
Gemini-1.0-Pro-Vis	62.90	-	57.63	-	40.17	-	53.57	-
GPT-4V	42.46	-	65.22	-	33.76	-	47.15	-
Likelihood Evaluation								
BLIP-2$_F$	48.78	0.05	61.36	0.11	43.21	0.13	51.12	0.10
InstructBLIP$_F$	48.29	0.08	60.51	0.17	44.82	0.12	51.21	0.12
InstructBLIP$_V$	53.19	0.06	59.15	0.19	44.40	0.16	52.25	0.14
LLaVA-1.0-7B$_V$	38.96	0.24	52.54	0.21	35.81	0.31	42.44	0.25
LLaVA-1.0-7B$_{L_2}$	45.73	0.22	59.66	0.16	36.66	0.22	47.35	0.20
MiniGPT4	49.37	0.39	57.12	0.17	41.18	0.21	49.22	0.26
mPLUG-Owl	46.14	0.18	59.15	0.17	40.59	0.22	48.63	0.19
PandaGPT	36.67	0.31	52.03	0.29	29.60	0.33	39.43	0.31
ImageBindLLM	43.39	0.20	54.07	0.16	40.89	0.20	46.12	0.19
LA-V2	42.92	0.14	60.85	0.15	42.16	0.18	48.64	0.16
mmGPT	49.91	0.15	50.85	0.23	40.85	0.20	47.20	0.19
Shikra	42.72	0.11	57.12	0.25	36.62	0.23	45.49	0.20
Cheetor$_V$	48.61	0.20	60.00	0.19	36.49	0.33	48.37	0.24
Cheetor$_{L_2}$	47.33	0.20	58.31	0.18	40.34	0.20	48.66	0.19
BLIVA	46.52	0.05	63.39	0.20	45.20	0.18	51.70	0.14
LLaVA-1.5-7B$_V$	51.01	0.00	65.25	0.00	43.55	0.00	53.27	0.00
MiniGPT-v2	56.38	0.00	65.25	0.00	45.67	0.00	55.77	0.00
Qwen-VL-Chat	45.94	0.00	61.86	0.00	43.34	0.00	50.38	0.00
LLaVA-1.6-7B$_V$	54.41	0.00	63.56	0.00	42.71	0.00	53.56	0.00
Monkey	44.52	0.00	65.25	0.00	44.18	0.00	51.32	0.00
Deepseek-VL	55.28	0.00	64.41	0.01	43.51	0.00	54.40	0.00
ShareGPT4V-7B	59.16	0.00	64.41	0.00	47.36	0.00	56.98	0.00
ShareGPT4V-13B	56.43	0.00	65.25	0.00	45.88	0.00	55.85	0.00
OmniLMM-12B	**77.22**	0.00	**70.34**	0.00	**51.37**	0.00	**66.31**	0.00
LLaVA-1.5-13B$_V$	50.26	0.00	68.64	0.00	47.57	0.00	55.49	0.00
LLaVA-1.6-13B$_V$	54.93	0.00	67.80	0.00	45.45	0.00	56.06	0.00
Qwen-VL-Max	-	-	-	-	-	-	-	-
Gemini-1.0-Pro-Vis	-	-	-	-	-	-	-	-
GPT-4V	-	-	-	-	-	-	-	-

Table 23: Evaluation results on spatial understanding.

Model	Image Captioning				Avg.
	COCO	TextCaps	NoCaps	Flickr30K	
BLIP-2$_F$	**97.48**	41.56	83.57	**74.63**	83.57
InstructBLIP$_F$	54.79	16.38	45.31	58.63	45.31
InstructBLIP$_V$	30.97	17.16	30.18	30.77	30.18
LLaVA-1.0-7B$_V$	47.16	21.79	42.43	35.78	42.43
LLaVA-1.0-7B$_{L_2}$	50.74	24.49	45.44	37.45	45.44
MiniGPT4	57.20	29.19	58.71	44.71	58.71
mPLUG-Owl	59.36	24.25	48.43	46.61	48.43
PandaGPT	2.24	0.95	1.12	1.93	1.12
ImageBindLLM	38.15	16.45	32.83	23.14	32.83
LA-V2	44.60	22.10	41.06	36.08	41.06
mmGPT	35.50	18.68	33.20	23.45	33.20
Shikra	41.01	19.76	37.42	28.91	37.42
Cheetor$_V$	86.90	32.70	73.99	52.88	73.99
Cheetor$_{L_2}$	72.80	21.64	44.39	36.63	44.39
BLIVA	62.23	36.72	64.21	46.90	64.21
LLaVA-1.5-7B$_V$	78.61	58.61	79.30	69.57	79.30
MiniGPT-v2	8.15	7.42	6.73	8.38	6.73
Qwen-VL-Chat	59.87	62.81	54.56	52.28	54.56
LLaVA-1.6-7B$_V$	57.15	34.89	52.57	47.22	52.57
Monkey	38.09	43.55	54.60	43.65	54.60
Deepseek-VL	65.91	57.23	66.33	62.32	66.33
ShareGPT4V-7B	78.72	64.19	84.17	73.68	84.17
ShareGPT4V-13B	82.22	**68.84**	**91.41**	71.82	**91.41**
OmniLMM-12B	54.90	49.34	58.42	73.05	58.42
LLaVA-1.5-13B$_V$	85.05	65.07	84.82	69.88	84.82
LLaVA-1.6-13B$_V$	57.95	33.47	50.54	43.98	50.54
Qwen-VL-Max	51.37	48.99	76.84	74.32	76.84
Gemini-1.0-ProV	49.08	12.25	52.77	58.46	52.77
GPT-4V	33.18	36.86	24.77	26.57	24.77

Table 24: Evaluation results on visual description based on CIDEr.

Model	Image Captioning											
	COCO			TextCaps			NoCaps			Flickr30K		
	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L
BLIP-2$_F$	**30.14**	26.71	49.57	16.84	18.95	34.26	37.18	27.09	49.34	23.63	23.37	44.24
InstructBLIP$_F$	14.24	18.85	34.98	2.74	10.63	18.58	15.59	17.60	32.94	19.07	20.51	38.34
InstructBLIP$_V$	5.91	13.09	23.96	3.30	10.51	18.19	7.71	13.48	25.13	6.26	12.09	24.90
LLaVA-1.0-7B$_V$	14.48	20.47	34.87	6.86	15.34	25.99	17.39	21.67	36.27	11.76	21.87	33.10
LLaVA-1.0-7B$_{L_2}$	14.85	21.26	37.98	8.25	16.32	28.33	18.31	22.50	39.15	13.93	22.22	35.90
MiniGPT4	18.31	22.47	37.65	9.66	17.34	29.22	22.84	24.99	40.83	13.82	22.28	34.48
mPLUG-Owl	18.30	21.55	40.19	7.32	15.56	27.21	17.35	21.68	37.66	16.57	23.34	39.93
PandaGPT	1.31	8.77	21.34	1.38	7.85	20.89	1.74	8.98	22.38	0.00	7.44	17.62
ImageBindLLM	11.27	18.76	32.30	5.14	13.79	24.66	12.53	18.79	32.56	6.91	17.10	27.73
LA-V2	13.06	19.78	33.51	6.82	15.47	25.59	15.45	21.32	35.61	10.93	21.82	32.29
mmGPT	9.08	16.89	29.89	4.66	13.74	23.89	10.90	18.29	31.44	7.61	18.19	27.91
Shikra	12.47	19.30	31.65	6.25	14.92	23.35	13.68	20.51	32.70	9.47	20.77	28.40
Cheetor$_V$	28.12	26.01	**50.55**	12.51	17.88	33.34	31.96	26.66	49.99	22.71	25.23	43.47
Cheetor$_{L_2}$	23.03	23.44	46.47	8.53	15.13	28.66	17.87	20.34	39.27	14.33	21.45	39.13
BLIVA	11.57	20.76	35.67	12.04	18.73	30.67	21.89	23.06	40.63	8.40	19.32	33.10
LLaVA-1.5-7B$_V$	23.94	24.77	47.35	17.18	21.48	38.23	34.32	28.15	52.56	24.90	26.72	47.09
MiniGPT-v2	2.90	8.94	20.37	3.10	9.15	19.79	3.76	9.58	21.71	4.40	9.30	18.98
Qwen-VL-Chat	17.49	24.12	39.59	17.89	23.09	38.34	22.09	25.20	41.18	16.64	25.20	38.45
LLaVA-1.6-7B$_V$	16.83	20.95	38.05	9.22	17.21	28.44	19.82	22.22	39.21	15.78	22.75	37.99
Monkey	4.00	14.01	27.18	12.49	18.25	30.45	19.09	20.28	37.87	9.33	19.01	36.60
Deepseek-VL	19.55	23.73	43.78	15.59	22.30	37.00	26.13	26.36	47.44	20.42	26.09	42.96
ShareGPT4V-7B	24.47	24.41	47.18	19.30	22.31	39.32	28.10	27.88	53.21	**28.10**	26.96	47.16
ShareGPT4V-13B	26.22	25.49	49.07	**21.34**	22.96	40.39	**38.41**	28.81	55.30	26.94	27.47	48.26
OmniLMM-12B	15.89	22.09	39.40	12.58	20.02	32.03	24.24	24.66	43.12	25.03	27.20	46.62
LLaVA-1.5-13B$_V$	26.98	25.88	49.22	19.35	22.79	**40.79**	36.73	28.53	**55.43**	26.06	28.06	47.88
LLaVA-1.6-13B$_V$	16.95	21.90	38.54	8.65	16.82	27.59	19.13	22.37	38.57	15.24	22.54	35.51
Qwen-VL-Max	17.48	**29.12**	45.64	9.64	**25.91**	38.11	31.12	**29.95**	52.15	25.59	**30.39**	**48.58**
Gemini-1.0-ProV	14.86	26.74	41.84	9.35	19.13	22.40	23.01	28.51	46.22	20.59	27.93	44.33
GPT-4V	8.45	17.08	28.85	8.10	17.85	27.12	11.95	25.66	37.04	7.12	17.29	26.86

Table 25: Evaluation results on visual description based on BLEU-4, METEOR and ROUGE-L.

Model	ITM								VE				Avg.	
	MSCOCO$_{itm}$		**MSCOCO$_{its}$**		**WikiHow**		**Winoground**		**SNLI-VE**		**MOCHEG**			
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
Generation Evaluation														
BLIP-2$_F$	96.40	0.04	97.53	0.02	33.67	0.26	55.25	0.32	72.20	0.15	46.33	0.02	51.86	0.19
InstructBLIP$_F$	**97.63**	0.02	98.27	0.01	46.87	0.20	64.50	0.25	**74.80**	0.11	46.09	0.24	58.07	0.20
InstructBLIP$_V$	63.73	0.24	96.33	0.04	33.13	0.17	55.00	0.51	32.33	0.00	42.07	0.19	40.63	0.22
LLaVA-1.0-7B$_V$	51.40	0.10	77.67	0.23	29.47	0.47	46.00	0.55	35.87	0.33	43.37	0.31	38.68	0.42
LLaVA-1.0-7B$_{L_2}$	52.17	0.01	87.20	0.11	34.00	0.29	49.00	0.42	33.13	0.07	44.02	0.56	40.04	0.34
MiniGPT4	52.20	0.00	61.87	0.26	27.87	0.36	53.00	0.57	37.33	0.29	40.36	0.65	39.64	0.47
mPLUG-Owl	51.63	0.19	42.00	0.68	26.20	0.79	49.50	0.56	36.60	0.62	36.33	0.65	37.16	0.66
PandaGPT	52.10	0.00	22.47	0.36	23.47	0.32	51.50	0.57	34.40	0.06	38.46	0.64	36.96	0.40
ImageBindLLM	51.87	0.13	29.20	0.43	21.53	0.54	48.75	0.54	32.07	0.54	35.92	0.66	34.57	0.57
LA-V2	52.20	0.00	42.00	0.73	25.93	0.83	48.00	0.54	37.00	0.69	41.24	0.14	38.04	0.55
mmGPT	51.73	0.29	32.00	0.87	24.47	0.86	50.25	0.59	32.33	0.59	38.34	0.57	36.35	0.65
Shikra	30.20	0.13	81.40	0.26	37.07	0.23	51.75	0.58	34.47	0.20	32.35	0.73	38.91	0.44
Cheetor$_V$	53.60	0.08	71.47	0.33	31.40	0.62	53.00	0.52	35.13	0.49	39.88	0.63	39.85	0.57
Cheetor$_{L_2}$	52.27	0.02	52.67	0.32	30.40	0.31	50.50	0.49	32.80	0.01	45.44	0.50	39.79	0.32
BLIVA	50.50	0.55	67.27	0.29	30.47	0.33	47.00	0.58	33.80	0.22	42.25	0.47	38.38	0.40
LLaVA-1.5-7B$_V$	73.50	0.42	86.60	0.31	43.27	0.56	66.25	0.32	38.33	0.41	45.44	0.46	48.32	0.44
MiniGPT-v2	53.70	0.49	31.60	0.81	23.93	0.79	43.75	0.54	32.67	0.42	40.30	0.62	35.16	0.59
Qwen-VL-Chat	84.00	0.27	87.20	0.29	39.13	0.53	65.25	0.22	51.73	0.39	42.01	0.33	49.53	0.37
LLaVA-1.6-7B$_V$	74.00	0.34	86.47	0.31	38.00	0.58	63.25	0.38	46.07	0.50	45.27	0.37	48.15	0.46
Monkey	79.70	0.35	96.47	0.06	49.90	0.50	64.75	0.37	46.73	0.41	40.41	0.66	50.45	0.48
Deepseek-VL	69.33	0.52	49.67	0.87	33.40	0.84	70.75	0.23	44.20	0.79	48.40	0.21	49.19	0.52
ShareGPT4V-7B	71.23	0.41	97.93	0.02	48.93	0.42	64.75	0.28	38.00	0.24	44.56	0.34	49.06	0.32
ShareGPT4V-13B	72.36	0.36	87.47	0.30	57.91	0.58	76.25	0.21	50.33	0.50	43.37	0.44	56.97	0.43
OmniLMM-12B	88.40	0.23	84.40	0.34	46.27	0.46	83.25	0.18	58.60	0.38	46.45	0.32	58.64	0.34
LLaVA-1.5-13B$_V$	70.97	0.37	87.67	0.30	43.53	0.56	69.50	0.29	49.07	0.52	46.21	0.34	52.08	0.43
LLaVA-1.6-13B$_V$	77.73	0.21	99.13	0.01	45.67	0.36	62.25	0.29	55.80	0.28	45.27	0.50	52.25	0.36
Qwen-VL-Max	97.32	-	**99.66**	-	57.77	-	**83.75**	-	64.00	-	50.63	-	64.04	-
Gemini-1.0-Pro-Vis	95.83	-	98.33	-	**73.33**	-	81.25	-	68.67	-	49.11	-	68.09	-
GPT-4V	94.66	-	98.00	-	71.67	-	**83.75**	-	71.00	-	**53.35**	-	**69.94**	-
Likelihood Evaluation														
BLIP-2$_F$	**96.37**	0.04	62.07	0.14	32.47	0.06	58.50	0.04	57.73	0.08	46.33	0.09	48.76	0.07
InstructBLIP$_F$	90.97	0.10	50.00	0.09	30.80	0.10	62.75	0.08	54.57	0.12	43.91	0.23	48.01	0.13
InstructBLIP$_V$	87.37	0.19	61.33	0.10	29.67	0.13	68.00	0.04	49.73	0.39	36.27	0.13	45.92	0.17
LLaVA-1.0-7B$_V$	48.30	0.01	72.40	0.09	30.47	0.17	63.75	0.06	39.13	0.39	33.96	0.16	41.83	0.20
LLaVA-1.0-7B$_{L_2}$	64.13	0.17	66.67	0.07	31.60	0.13	58.00	0.03	37.60	0.07	39.88	0.21	41.77	0.11
MiniGPT4	78.27	0.18	60.13	0.10	30.27	0.11	65.00	0.01	40.73	0.47	31.18	0.41	41.8	0.25
mPLUG-Owl	53.50	0.02	68.53	0.07	31.13	0.09	65.75	0.03	36.87	0.12	42.78	0.34	44.13	0.15
PandaGPT	49.13	0.47	26.27	0.15	26.40	0.21	47.25	0.13	33.80	0.52	38.88	0.40	36.58	0.32
ImageBindLLM	52.13	0.00	61.87	0.07	29.53	0.11	55.00	0.03	33.60	0.03	41.42	0.04	39.89	0.05
LA-V2	64.50	0.26	60.20	0.07	29.80	0.11	64.00	0.03	39.27	0.42	41.36	0.02	43.61	0.15
mmGPT	52.17	0.00	51.93	0.08	28.53	0.09	58.25	0.10	32.33	0.00	41.54	0.00	40.16	0.05
Shikra	90.63	0.14	78.13	0.08	31.87	0.08	64.25	0.03	49.40	0.10	41.36	0.00	46.72	0.05
Cheetor$_V$	79.07	0.26	58.07	0.17	29.93	0.21	62.00	0.05	40.67	0.54	34.08	0.10	41.67	0.23
Cheetor$_{L_2}$	56.13	0.08	63.33	0.10	29.80	0.16	58.50	0.06	34.40	0.05	41.12	0.17	40.96	0.11
BLIVA	83.30	0.27	59.33	0.14	31.40	0.12	63.75	0.10	42.40	0.19	42.25	0.25	44.95	0.17
LLaVA-1.5-7B$_V$	71.40	0.22	57.93	0.06	32.13	0.10	63.50	0.02	49.93	0.07	44.08	0.24	47.41	0.11
MiniGPT-v2	52.43	0.07	39.40	0.20	28.33	0.19	55.75	0.09	33.73	0.01	28.99	0.25	36.70	0.13
Qwen-VL-Chat	95.33	0.04	**80.93**	0.06	35.40	0.18	72.75	0.08	**60.00**	0.09	34.62	0.26	50.69	0.15
LLaVA-1.6-7B$_V$	69.17	0.19	55.07	0.07	31.73	0.10	64.75	0.05	50.33	0.13	43.37	0.24	47.55	0.13
Monkey	94.83	0.08	74.67	0.07	33.60	0.10	71.25	0.03	54.27	0.52	43.79	0.31	**50.73**	0.24
Deepseek-VL	67.97	0.06	43.87	0.12	28.60	0.17	60.50	0.07	44.20	0.16	**47.04**	0.24	45.09	0.16
ShareGPT4V-7B	86.13	0.27	71.13	0.07	32.73	0.07	70.50	0.02	54.53	0.12	43.37	0.20	50.28	0.10
ShareGPT4V-13B	77.33	0.09	30.93	0.05	31.93	0.12	**77.00**	0.01	50.33	0.12	42.96	0.27	50.56	0.13
OmniLMM-12B	83.67	0.01	62.93	0.06	31.33	0.14	71.00	0.05	53.33	0.06	24.50	0.05	45.04	0.07
LLaVA-1.5-13B$_V$	78.90	0.08	60.73	0.04	32.00	0.09	75.50	0.03	52.00	0.13	23.96	0.24	45.87	0.12
LLaVA-1.6-13B$_V$	84.30	0.08	72.73	0.07	**35.80**	0.09	65.75	0.04	56.60	0.16	41.07	0.15	49.81	0.11
Qwen-VL-Max	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Gemini-1.0-Pro-Vis	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4V	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 26: Evaluation results on cross-modal inference.

Model	VQA-MT		VisDial		Avg.	
	Acc	Instability	Acc	Instability	Acc	Instability
Generation Evaluation						
BLIP-2$_F$	67.97	0.20	55.53	0.24	55.53	0.24
InstructBLIP$_F$	68.67	0.19	52.51	0.24	52.51	0.24
InstructBLIP$_V$	56.58	0.48	40.68	0.61	40.68	0.61
LLaVA-1.0-7B$_V$	40.06	0.76	31.15	0.77	31.15	0.77
LLaVA-1.0-7B$_{L_2}$	50.47	0.54	42.11	0.46	42.11	0.46
MiniGPT4	43.98	0.23	35.05	0.66	35.05	0.66
mPLUG-Owl	38.66	0.77	31.85	0.80	31.85	0.80
PandaGPT	34.73	0.64	33.44	0.63	33.44	0.63
ImageBindLLM	37.24	0.66	33.26	0.69	33.26	0.69
LA-V2	38.88	0.72	32.00	0.76	32.00	0.76
mmGPT	34.92	0.80	28.75	0.90	28.75	0.90
Shikra	43.33	0.67	27.12	0.91	27.12	0.91
Cheetor$_V$	44.40	0.55	36.14	0.59	36.14	0.59
Cheetor$_{L_2}$	41.36	0.49	39.80	0.39	39.80	0.39
BLIVA	48.83	0.57	30.80	0.75	30.80	0.75
LLaVA-1.5-7B$_V$	67.27	0.28	56.91	0.39	56.91	0.39
MiniGPT-v2	37.54	0.52	37.54	0.45	37.54	0.45
Qwen-VL-Chat	72.20	0.28	55.60	0.49	55.60	0.49
LLaVA-1.6-7B$_V$	72.40	0.23	59.80	0.40	59.80	0.40
Monkey	70.89	0.31	48.80	0.58	48.80	0.58
Deepseek-VL	79.71	0.14	71.16	0.23	71.16	0.23
ShareGPT4V-7B	72.08	0.20	60.80	0.32	60.80	0.32
ShareGPT4V-13B	78.53	0.16	67.54	0.24	67.54	0.24
OmniLMM-12B	**82.37**	0.12	77.84	0.21	77.84	0.21
LLaVA-1.5-13B$_V$	67.27	0.02	66.70	0.25	66.70	0.25
LLaVA-1.6-13B$_V$	78.97	0.13	69.26	0.24	69.26	0.24
Qwen-VL-Max	**85.54**	-	**81.50**	-	**81.50**	-
Gemini-1.0-Pro-Vis	80.45	-	71.50	-	71.50	-
GPT-4V	52.29	-	76.60	-	76.60	-
Likelihood Evaluation						
BLIP-2$_F$	71.52	0.04	53.62	0.04	53.62	0.04
InstructBLIP$_F$	77.06	0.06	57.34	0.04	57.34	0.04
InstructBLIP$_V$	78.06	0.04	59.30	0.04	59.30	0.04
LLaVA-1.0-7B$_V$	61.32	0.05	43.24	0.04	43.24	0.04
LLaVA-1.0-7B$_{L_2}$	56.24	0.06	40.97	0.03	40.97	0.03
MiniGPT4	63.97	0.06	44.14	0.05	44.14	0.05
mPLUG-Owl	52.38	0.04	38.57	0.03	38.57	0.03
PandaGPT	43.71	0.18	39.21	0.09	39.21	0.09
ImageBindLLM	43.11	0.03	35.86	0.02	35.86	0.02
LA-V2	47.29	0.08	39.49	0.04	39.49	0.04
mmGPT	47.52	0.06	38.57	0.03	38.57	0.03
Shikra	69.15	0.06	49.76	0.03	49.76	0.03
Cheetor$_V$	66.01	0.06	49.22	0.06	49.22	0.06
Cheetor$_{L_2}$	51.86	0.10	41.82	0.05	41.82	0.05
BLIVA	77.92	0.05	58.42	0.04	58.42	0.04
LLaVA-1.5-7B$_V$	76.79	0.02	59.40	0.03	59.40	0.03
MiniGPT-v2	46.21	0.02	38.89	0.06	38.89	0.06
Qwen-VL-Chat	81.73	0.03	60.96	0.04	60.96	0.04
LLaVA-1.6-7B$_V$	77.00	0.02	58.50	0.02	58.50	0.02
Monkey	81.57	0.03	50.18	0.04	50.18	0.04
Deepseek-VL	**84.87**	0.02	63.52	0.04	63.52	0.04
ShareGPT4V-7B	79.97	0.02	60.16	0.01	60.16	0.01
ShareGPT4V-13B	80.73	0.01	61.26	0.02	61.26	0.02
OmniLMM-12B	84.01	0.02	**65.18**	0.03	**65.18**	0.03
LLaVA-1.5-13B$_V$	78.82	0.01	58.48	0.02	58.48	0.02
LLaVA-1.6-13B$_V$	77.43	0.02	59.66	0.02	59.66	0.02
Qwen-VL-Max	-	-	-	-	-	-
Gemini-1.0-Pro-Vis	-	-	-	-	-	-
GPT-4V	-	-	-	-	-	-

Table 27: Evaluation results on multi-turn Dialogue. "Corr" represents the correlation coefficient between the model performance and the number of dialogue turns.