LENS: MULTI-LEVEL EVALUATION OF MULTIMODAL REASONING WITH LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have achieved significant advances in integrating visual and linguistic information, yet their ability to reason about complex and real-world scenarios remains limited. Existing benchmarks are usually constructed in a task-oriented manner, without a guarantee that different task samples come from the same data distribution. Therefore, they often fall short in evaluating the synergistic effects of lower-level perceptual capabilities on higher-order reasoning. To lift this limitation, we contribute `Lens`, a multi-level evaluation benchmark of multimodal reasoning with 3.4K contemporary images and 60K+ human-authored questions covering eight tasks and 12 daily scenarios, forming three progressive task tiers, *i.e.*, perception, understanding, and reasoning. One feature is that each image is equipped with rich annotations for all tasks. Thus, this data set intrinsically supports evaluating MLLMs to handle image-invariable prompts, from basic perception to compositional reasoning. In addition, our images have been collected manually from social media, with 53% published after Jan. 2025. We evaluate 15+ frontier MLLMs such as Qwen2.5-VL, InternVL3, GPT-4o and two reasoning models QVQ-Max and Kimi-VL. Most models were released in 2025, and none of them achieve an accuracy beyond 60% in the reasoning tasks. Furthermore, we propose the Self-Driven Multi-Expert Collaborative Framework (SMEC), a framework designed for MLLMs that simulates a panel of experts discussing and exchanging viewpoints via self-generated role-specific prompts. The experimental results confirm the existence of synergistic effects in a hierarchical task structure, where low-level tasks facilitate the reasoning of MLLMs on more complex, high-level tasks. Statistical analysis and ablation studies further demonstrate the comprehensiveness of our dataset and the superiority of our methodology. Our dataset and code will be released upon acceptance.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have emerged as a rapidly advancing field in artificial intelligence, demonstrating substantial improvements in visual content recognition and multimodal reasoning (Zhu et al., 2025; Bai et al., 2025; Wu et al., 2024; Team et al., 2024; 2025a). Despite their promising capabilities, MLLMs continue to face significant challenges in interpreting complex and real-world visual environments that are inherently dynamic, diverse, and grounded in physicality. However, existing benchmarks remain limited in their ability to evaluate multi-level reasoning.

Early evaluations were largely based on classical computer vision tasks (Everingham et al., 2010; Lin et al., 2014; Yu et al., 2016) and their integration with natural language. The real-world knowledge was often superficial, resulting in weak alignment between visual input and linguistic output. Secondly, these benchmarks are typically constructed under closed-world assumptions, lacking the inter-task consistency needed to assess reasoning across modalities (Fu et al., 2024; Li et al., 2024a). As a result, the absence of quantitative multi-level evaluation hinders meaningful comparison across MLLMs.

More recent benchmarks have begun to shift toward open-world evaluation and multimodal reasoning tasks (Wu & Xie, 2024; Zang et al., 2025). While this represents progress, current benchmarks do not adequately assess the nuanced performance necessary to evaluate MLLMs' progression towards human-like intelligence in real-world settings. They require largely primary visual comprehension and fall short of measuring higher-order reasoning and spatial understanding (Yue et al., 2024; Liu et al., 2024c). Furthermore, data distributions often differed between tasks, so that high performance in perceptual tasks did not necessarily translate into strong inference capabilities in more complex integrated multimodal tasks (Fan et al., 2025). As a result, they ignore the synergistic effect of the combinations of lower-order perceptual abilities on higher-order reasoning and are hard to provide a fine-grained assessment.

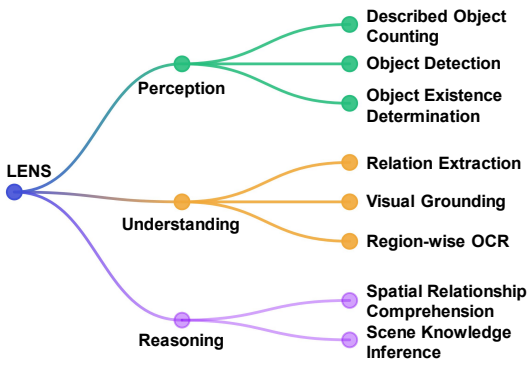


Figure 1: Illustration of the task split in Lens.

In this study, we propose a hierarchical and comprehensive evaluation framework Lens specifically designed to assess the multimodal capabilities in real-world scenarios. Our benchmark focuses on both isolated tasks and the integration of perception, understanding, and reasoning—three core tiers essential for intelligent multimodal systems. As shown in Figure 1, Lens encompasses eight tasks, systematically organized into three hierarchical tiers with eight subtasks, and it comprises 3.4K real-world photographs and 60K+ human-authored questions, in 12 diverse scenarios—including streets, stations, schools, homes, and more, which can be roughly divided into three themes: "Home", "Education", and "City", and we visualize the high-frequency words under different themes in Figure 2. 53% of the images are from 2025 and more than 80% of the images are from after September 2024, ensuring the content reflects contemporary environments.

For task design, Lens adopts an open-set configuration, allowing queries to be posed in natural language and grounded in authentic photographic content. This design enables evaluation of model performance in complex, ambiguous, and information-rich settings, better aligning with real-time human demands. Moreover, our benchmark introduces multi-level tasks, which are unified by shared visual contexts, making Lens well-suited for assessing the synergistic effects of lower-level perceptual abilities (*e.g.*, object detection, localization) on higher-order reasoning tasks. To succeed in Lens, models must jointly process multimodal input, recall domain knowledge, and conduct multi-step reasoning to arrive at valid conclusions. Our experimental results confirm that current state-of-the-art MLLMs still struggle with these reasoning-heavy tasks, revealing a significant gap between perception and functional understanding.

To bridge this gap, we propose the Self-Driven Multi-Expert Collaborative Framework (SMEC), a novel reasoning framework that leverages the MLLM itself as a set of specialized experts instantiated through self-generated prompts. Unlike tool-calling approaches (Wang et al., 2025; Gao et al., 2025; Liu et al., 2025b; Zhang et al., 2024b) that rely on external modules, SMEC treats the base MLLM as a versatile reasoning engine: it simulates diverse expert perspectives (*e.g.*, spatial analyst, text interpreter, commonsense reasoner) via role-specific prompts and composes their insights into coherent final answers. This collaborative mechanism encourages the model to extract, expand, and integrate rich, task-relevant information. Our experiments demonstrate that SMEC significantly boosts performance on reasoning tasks within Lens, validating its potential as a general-purpose, language-native method for enhancing multimodal reasoning.

In sum, we make the following contributions:

- **Realistic and Up-to-Date Evaluation.** By leveraging a newly collected set of high-resolution, naturalistic images, our benchmark evaluates the latest multimodal reasoning models in settings that closely reflect real-world complexity.

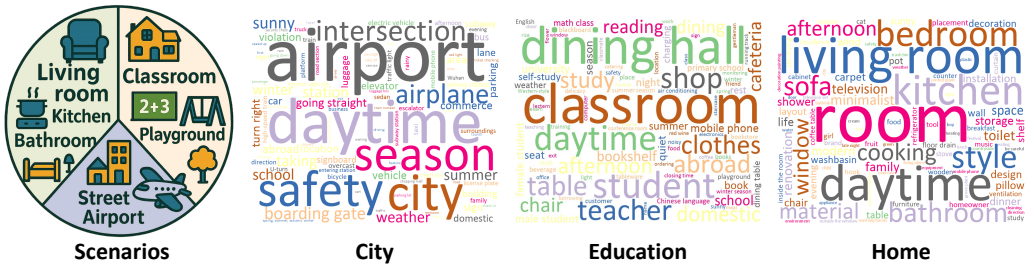


Figure 2: Three core themes, "Education", "City", and "Home", along with their word clouds of the scenario distributions by name size.

- **Multi-Level Evaluation.** It supports fine-grained and interpretable evaluation across three core dimensions—perception, understanding, and reasoning—providing a comprehensive view of a model's multimodal competence.
- **Synergistic Capability Evaluation.** Unlike existing benchmarks that often assess tasks in isolation, our framework emphasizes the synergistic effects of lower-level perceptual abilities on higher-order reasoning tasks. The experimental results also confirm that low-level tasks facilitate the reasoning of MLLMs on more complex, high-level tasks (*e.g.*, Scene Knowledge Inference).
- **Towards Generalizable Intelligence.** By capturing both perceptual and reasoning performance in integrated tasks, our benchmark helps identify the gaps between current model capabilities and the requirements of human-aligned reasoning systems and measure the shortcomings of current models.
- **Self-driven Reasoning Enhancement.** We introduce SMEC, a self-driven multi-expert collaborative framework that simulates specialized experts within a single MLLM through self-generated prompts. Unlike tool-calling approaches, SMEC enables modular, multi-perspective reasoning natively, leading to significant gains on complex reasoning tasks.

Comparison with existing benchmarks. Compared with existing multimodal benchmarks (Liu et al., 2024c; Yue et al., 2024; Li et al., 2024b), `Lens` provides more contemporary, diverse, and densely annotated visual content. Our benchmark is constructed from contemporary social-media images, ensuring strong timeliness and significantly reducing the risk of contamination from pre-training corpora. In contrast to task-specific datasets (Liu et al., 2024d;a; Wei et al., 2024), our benchmark provides rich, multi-task annotations with the same visual content, across perception, understanding, and reasoning, enabling controlled analysis of cross-task synergies within a unified distribution. Additionally, `Lens` offers the detailed thought process in real-world reasoning tasks for potential future research. Appendix A.1 further discusses related work.

2 LENS DATASET AND BENCHMARK

2.1 DATA COLLECTION

The image data collection in our benchmark focuses on real-world scenes to ensure diversity, representativeness, and practicality for visual perception, understanding, and reasoning tasks. To this end, we first defined a set of common real-life scenarios that are highly relevant to typical human visual experiences. The selection principle was that each visual scene should contain distinguishable and representative semantic content. For example, street scenes are usually populated with cars, pedestrians, and storefronts, while indoor environments like classrooms often involve students, teachers, and educational materials. To avoid regional or cultural bias and ensure a broad distribution of content, we collected images from multiple social media platforms, including X (formerly Twitter), Instagram, Weibo, and RedNote. These platforms were chosen due to their global user bases and diverse content coverage across regions and lifestyles. During the collection process, we strictly complied with the copyright and licensing regulations of each platform, ensuring that data was collected only from publicly accessible posts and that no images were downloaded from sources

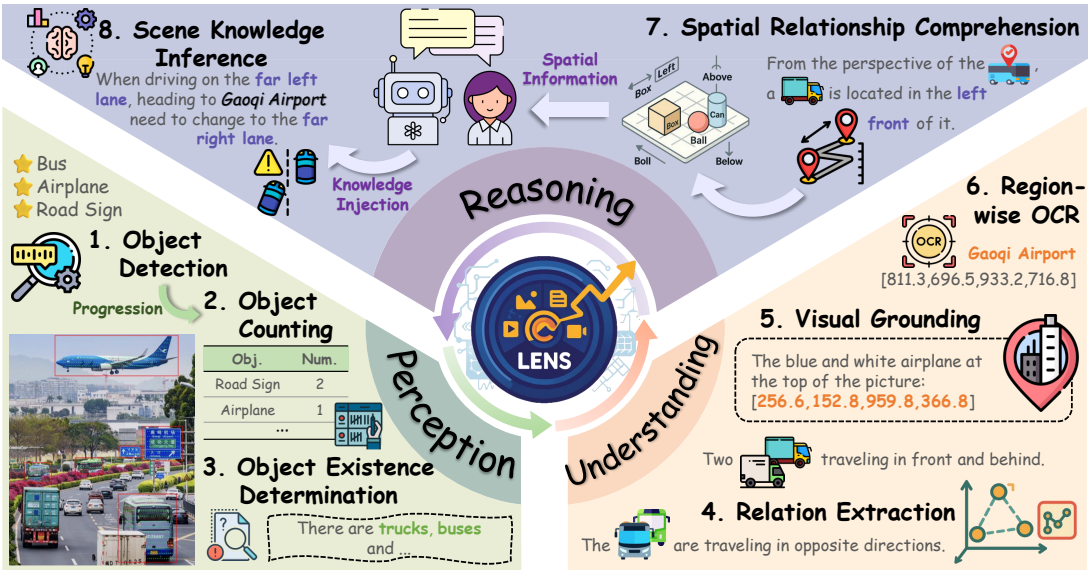


Figure 3: Lens consists of eight tasks at three levels. **Perception** tasks focus on recognizing object attribute and counting. **Understanding** tasks emphasizes localization and inter-object relationships with textual information. **Reasoning** tasks demand the use of external knowledge beyond the visual input and involve multi-step, complex reasoning processes to arrive at the correct answer.

Table 1: Comparison with other recently released multimodal benchmarks.

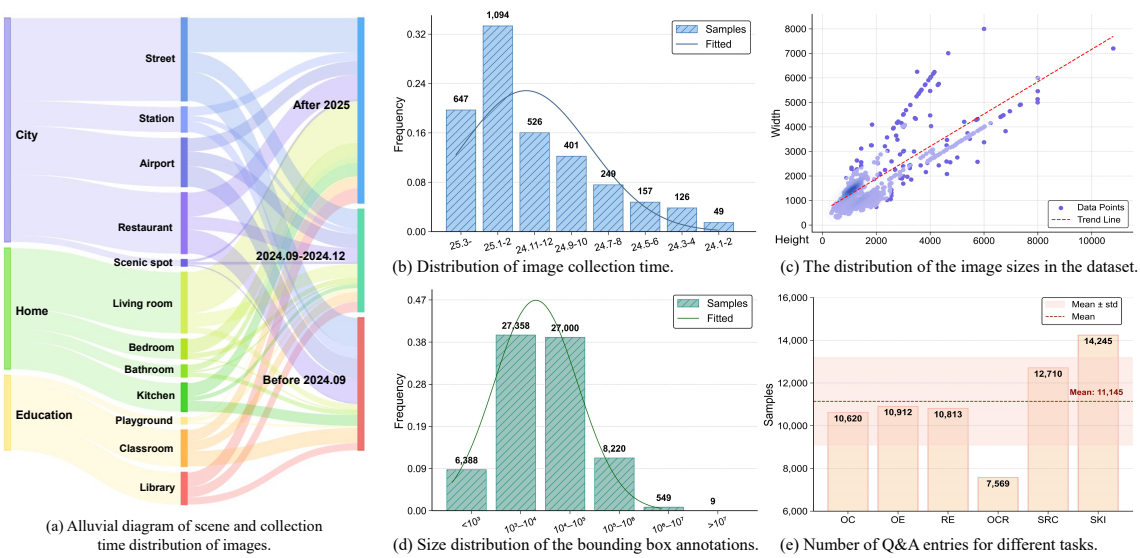
Benchmarks	Venue	Att.	Cnt	Loc	Rel	Reasoning	Interleaved Image-Text	Image Source
V* (Wu & Xie, 2024)	CVPR'24	✘	✘	✔	✔	✔	✘	SA-1B (Kirillov et al., 2023)
SPEC (Peng et al., 2024)	CVPR'24	✔	✔	✔	✘	✘	✘	Synthesize
MMVP (Tong et al., 2024)	CVPR'24	✔	✘	✘	✘	✘	✘	ImageNet (Russakovsky et al., 2015), LAION-5B (Schuhmann et al., 2022)
HaloQuest (Wang et al., 2024b)	ECCV'24	✔	✘	✘	✔	✔	✘	Open Images (Kuznetsova et al., 2020)
AS-V2 (Wang et al., 2024a)	ECCV'24	✔	✔	✔	✘	✔	✘	COCO (Caesar et al., 2018)
MMBench (Liu et al., 2024c)	ECCV'24	✔	✔	✔	✔	✔	✘	Internet images
HC-RefLoCo (Wei et al., 2024)	NeurIPS'24	✘	✘	✔	✔	✔	✘	Multiple existing datasets
Visual CoT (Shao et al., 2024)	NeurIPS'24	✔	✘	✘	✘	✔	✘	Multiple existing datasets
MC-Bench (Xu et al., 2024)	arXiv'24	✘	✘	✔	✘	✔	✔	Multiple existing datasets, Internet
CODE (Zang et al., 2025)	IJCV'25	✔	✔	✔	✘	✘	✘	Flickr30k series (Young et al., 2014; Plummer et al., 2015)
ChatterBox (Tian et al., 2025)	AAAI'25	✔	✔	✔	✘	✔	✘	Visual Genome (Krishna et al., 2017)
Lens	-	✔	✔	✔	✔	✔	✔	Collect manually from social media 53% published later than Jan. 2025

"Att.": Attribute; "Cnt": Count; "Loc": Localization; "Rel": Relation

explicitly prohibiting data reuse or redistribution. Moreover, to facilitate the evaluation of multiple subtasks within the same image (*e.g.*, detection, OCR, scene knowledge inference), we curated images that exhibit rich semantic content while maintaining scene clarity. Complex or ambiguous images were manually filtered out to avoid introducing noise that could hinder benchmarking or evaluation consistency. Please note that we manually collect these data that are completely open to the Internet and have complied with the developer agreement of the relevant platform (*e.g.*, Developer Policy of X [1] and Meta [2]), ensuring non-commercial use, erasing geographic information, user personal information, etc. from the original data. Images containing sensitive personal information were either excluded or processed to blur or mask sensitive regions, to mitigate privacy risks. For further details, please refer to the appendix A.4.

[1]https://developer.x.com/en/use-cases/do-research/academic-research

[2]https://developers.facebook.com/docs/instagram-platform



(a) Alluvial diagram of scene and collection time distribution of images.

(b) Distribution of image collection time.

(c) The distribution of the image sizes in the dataset.

(d) Size distribution of the bounding box annotations.

(e) Number of Q&A entries for different tasks.

Figure 4: Statistical analysis of our dataset. We visualize the temporal distribution of the images for different scenarios, size distribution of images and bounding box annotations, and number of QA entries for different tasks, demonstrating the timeliness and diversity of our data.

2.2 Task Design and Annotation Process

To construct a comprehensive and diverse benchmark, we recruited over 50 undergraduate and graduate students (including authors) as human annotators to assist in the process of question collection and task annotation and paid the corresponding salary. These annotators were carefully trained to ensure high annotation quality and consistency. As shown in Figure 3, the generated questions were divided into three major categories: Perception, Understanding, and Reasoning. For Perception and Understanding, they primarily target the model's ability to perceive visual objects and align them accurately with natural language descriptions. They emphasize fine-grained visual grounding and object recognition rather than abstract reasoning. At last, reasoning-based questions aim to evaluate the model's ability to understand user intent and reason based on external knowledge, commonsense, physical laws, or background information beyond the purely visual content of the image. Based on these assessment dimensions, we compare `Lens` with related multimodal benchmarks in Table 1 and formulate our challenging open-ended, language-driven tasks: Object Counting (OC), Object Detection (OD), Object Existence Determination (OE), Relation Extraction (RE), Visual Grounding (VG), Region-wise OCR (OCR), Spatial Relationship Comprehension (SRC), and Scene Knowledge Inference (SKI). We provide a more detailed introduction of these tasks in Appendix A.2.

2.3 Data Analysis

We aim to construct a dataset that is not only comprehensive and dynamic but also emphasizes reasoning ability practices. In the following analysis, we demonstrate the strengths of our benchmark in terms of diversity of images and annotations. The quantitative results are visualized in Figure 4.

First, our benchmark incorporates scene-aware content and real-time data. As shown in Figure 4, more than 50% of the images in our dataset were collected in 2025, and approximately 70% were collected in November 2024 and beyond, which avoids potential data leakage. Many images reflect dynamic scenes (*e.g.*, crowded streets, interactive environments) captured at different times and locations, aligning with real-world scenarios.

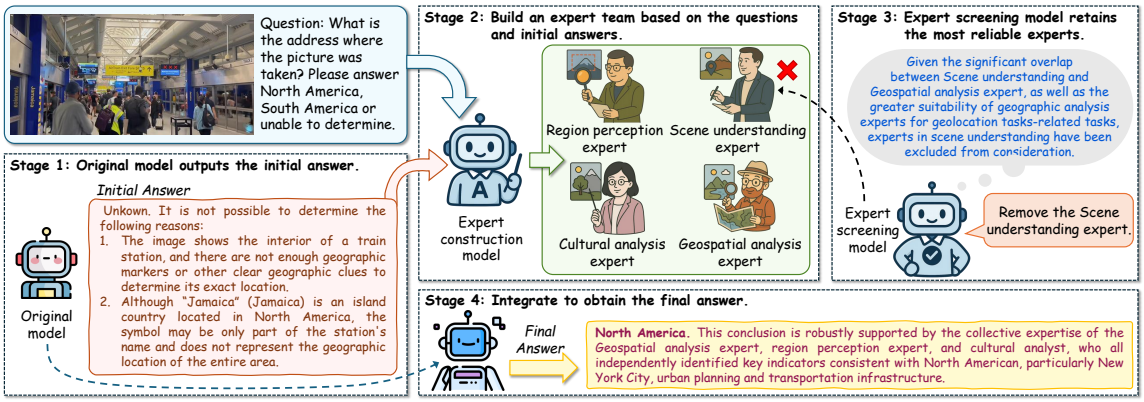


Figure 5: SMEC is a step-wise reasoning framework for answering complex visual questions using the Multi-Expert Collaboration. Starting from an initial model response, SMEC constructs a specialized expert team, to re-evaluate and refine the answer. Through expert screening and integration, unreliable and duplicated experts are filtered out, and a consensus-based final answer is produced, demonstrating the advantages of modular, expert-driven collaboration in visual reasoning tasks.

Second, in our dataset, the coverage of a wide range of object categories, scene types, and bounding box annotations further supports diverse downstream tasks from detection to high-level semantic inference and interleaved image-text understanding. As illustrated in Figure 4 (c), the high resolution of the images in our dataset makes it challenging for fine-grained understanding of the model and supports evaluation across varying input sizes. Additionally, as shown in Figure 4 (d), the various objects are labeled with different sizes of bounding boxes to meet the needs of multi-scale object detection and region-wise OCR evaluation.

Furthermore, beyond perception, our dataset facilitates reasoning-oriented research by supporting tasks that require: Spatial reasoning (*e.g.*, understanding object layouts and geometric relationships). Relational inference (*e.g.*, extracting interactions between objects). Commonsense knowledge application (*e.g.*, inferring the feasibility of a behavior or scene functionalities). Cross-modal alignment (*e.g.*, grounding free-form language to specific visual content). We also analyze the question-answer pairs distribution of these tasks and Figure 4 (e) shows that over 60% of the questions in the dataset go beyond simple recognition, explicitly encouraging models to reason about the scene, context, and user intent. Please refer to the Appendix A.5 for more low-level visual analysis.

3 Self-Driven Multi-Expert Collaborative Framework

We propose a Self-Driven Multi-Expert Collaborative Framework to tackle complex visual reasoning tasks that require diverse domain expertise and multi-level inference. Built on an instruction-tuned MLLM, our framework dynamically assembles a set of self-generated experts, each embodying a distinct reasoning perspective, as shown in Figure 5.

Expert Generation via Prompted Role Construction. Given a query q, the base model θ first produces a coarse initial answer a_0. To enrich the reasoning space, a Meta Generation Prompt p_g is used to iteratively generate expert role descriptions d_q^t, simulating specialized agents (*e.g.*, geospatial analyst, cultural analyst). Each valid description yields a new expert response a_t, which is added to the answer set A. This loop continues until either a diversity criterion is violated or a maximum number of iterations N_t is reached.

Prompt Adaptation and Redundancy Filtering. To avoid degenerate expert generation, the framework dynamically updates p_g when semantically redundant descriptions emerge. This adaptation encourages

exploration of novel expert roles. Implicit expert screening is performed by discarding repetitive or low-information descriptions, ensuring a concise yet diverse expert team with minimal computational overhead.

Consensus-Driven Answer Integration. The final stage aggregates the expert responses via a Collaboration Prompt p_c, prompting θ to synthesize a unified answer a_{final} through deliberative reasoning. This mimics human expert panels that reconcile differing views to reach a robust consensus.

We detail a formal description of the process in Appendix A.10. By this way, our framework instantiates modular experts purely via prompt-based self-conditioning. Unlike fixed-rule multi-agent systems and tool-calling methods, our framework leverages the generative flexibility of MLLM to dynamically instantiate and evolve its behaviors, without requiring external task-specific supervision.

4 EXPERIMENTS

4.1 EVALUATION MODELS

To illustrate the difficulty of our benchmark and evaluate the latest advances in current research, we evaluate various MLLMs belonging to three major categories: Closed-source generalist MLLMs, such as GPT-4o (Achiam et al., 2023) and Gemini2.5 Pro (Team et al., 2024). Open-source generalist MLLMs like Qwen2.5-VL (Bai et al., 2025), Deepseek-VL2 (Wu et al., 2024), Gemma3 (Team et al., 2025a), InternVL3 (Zhu et al., 2025). Multimodal reasoning models QvQ-Max, Kimi-VL-thinking (Team et al., 2025b) and GLM-4.1V-Thinking (Hong et al., 2025), focusing on advanced reasoning capabilities. The release dates of these models are distributed from Dec. 2024 to Apr. 2025.

4.2 EVALUATION STRATEGY

To ensure a fair and efficient assessment of model performance across our benchmark, we adopt two evaluation strategies for main results. For perception and understanding tasks, models were evaluated based on their direct outputs without additional inference-time computations. For complex reasoning tasks, which require deeper multi-step inference, we allow models to generate multiple candidate responses per question and the final prediction is then selected via majority voting (Liu et al., 2025a). For qualitative judgment, we follow prior work (Wang et al., 2023) and employ a large language model (*e.g.*, GLM4-flash (GLM et al., 2024)) as an automatic evaluator. The LLM is prompted to produce multiple pieces of evaluation evidence for calibration, comparing the model-generated responses against human-annotated answers, aiming to offer a consistent framework for evaluating model performance across diverse tasks.

4.3 EVALUATION RESULTS

We evaluate a suite of state-of-the-art Multimodal Large Language Models on our benchmark, which spans three tiers and eight tasks. Results, as shown in Table 2, reveal insights into model scaling, inter-task dependencies, and capability gaps in current MLLMs.

Model Scaling and General Trends. We observe a consistent performance gain with increased model size in both closed- and open-source models. For example, Qwen2.5-VL improves steadily from 3B to 72B, achieving top performance on reasoning tasks. InternVL3 shows similar gains in OD, rising from 18.39% (2B) to 47.44% (78B), though performance saturates at higher scales. These trends confirm that scaling remains a key driver for multimodal reasoning, albeit with diminishing returns in some subtasks.

Perception: Foundation for Higher Cognition. Perception-level tasks form the backbone of visual reasoning. Closed-source models like Gemini2.5-Pro and GPT-4o excel at OE (86.59% and 85.09%, respectively), although OD support is lacking. Among open-source models, Deepseek-VL2 and Qwen2.5-VL-72B deliver

Table 2: Comparison of state-of-the-art methods on `Lens`. We evaluate object detection (OD) performance using AP_{50} (Lin et al., 2014), visual grounding (VG) performance with ACC@0.5 (Xiao et al., 2024), and use accuracy for other tasks. Task abbreviations follow the definitions provided in Section 2.2. "MoE 1B/3B" denotes 3B Mixture of Experts model with 1B parameters activated. "N/A" denotes the official documentation does not confirm that the model is applicable for the task. Best performing models are shaded in red.

Methods	Model size	Perception			Understanding			Reasoning	
		OC	OD	OE	RE	VG	OCR	SRC	SKI
MLLM (closed source)									
GPT-4o	-	54.32	N/A	85.09	72.77	N/A	42.86	51.14	55.20
Gemini2.5-Pro	-	60.18	47.40	86.59	76.52	25.61	61.95	56.20	59.31
Open source									
Deepseek-VL2-tiny	MoE 1B/3B	56.22	21.12	72.11	58.73	16.09	44.01	38.97	45.12
Deepseek-VL2	MoE 4.5B/27B	61.41	46.08	77.68	69.18	42.47	48.76	44.58	49.50
Gemma3	4B	38.85	N/A	71.88	62.98	N/A	27.03	39.53	45.18
Gemma3	12B	44.65	N/A	73.21	62.78	N/A	33.98	43.33	48.56
InternVL3	2B	55.81	18.39	71.96	64.49	15.22	45.51	40.56	48.59
InternVL3	9B	55.63	25.79	77.49	67.18	18.18	48.79	44.69	51.32
InternVL3	38B	62.78	43.44	81.60	71.37	24.98	51.72	47.18	51.85
InternVL3	78B	61.38	47.44	84.87	74.93	27.24	54.21	49.39	55.17
Qwen2.5-VL	3B	58.76	35.16	74.01	66.52	39.44	52.43	40.33	46.50
Qwen2.5-VL	7B	58.35	37.75	83.75	71.58	40.11	61.65	46.28	48.87
Qwen2.5-VL	32B	62.25	39.93	83.60	74.57	41.15	65.64	51.66	51.54
Qwen2.5-VL	72B	59.75	43.48	85.67	75.98	44.98	68.51	53.65	54.79
Reasoning model									
QVQ-Max	72B	49.95	N/A	85.37	74.01	N/A	58.67	50.80	58.86
Kimi-VL-thinking	MoE 2.8B/16B	46.87	N/A	72.77	48.16	N/A	30.21	29.40	36.44

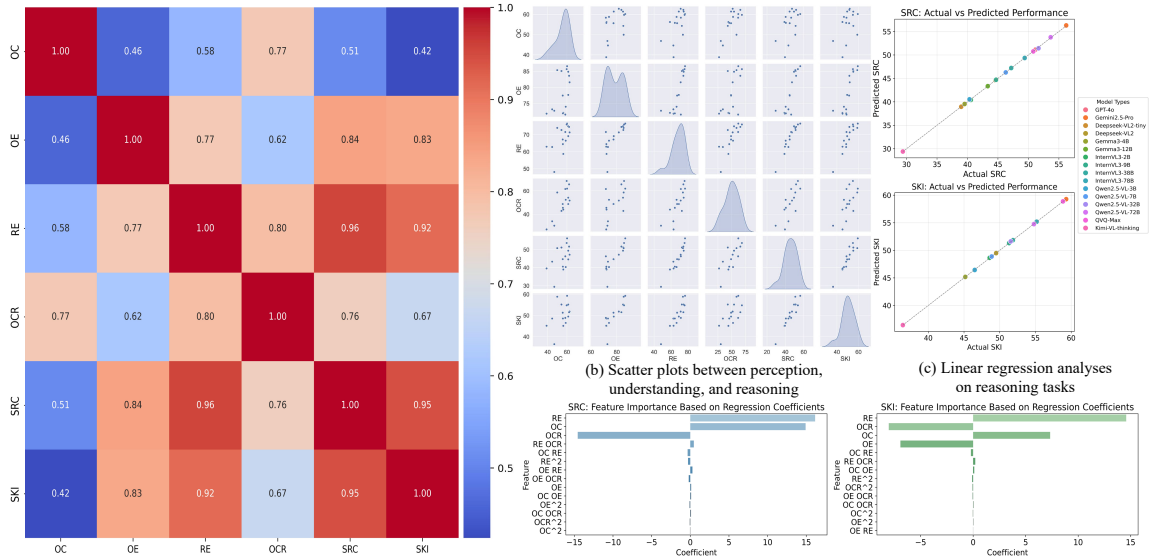
competitive OD and OC performance. Notably, models with stronger perception capabilities tend to exhibit superior reasoning performance, highlighting the foundational role of low-level visual understanding.

Understanding: Progress and Bottlenecks. Understanding tasks assess models' ability to interpret structured visual semantics with textual information. Gemini2.5-Pro leads in RE (76.52%) and OCR (61.95%), showcasing robust relational and textual grounding. However, VG remains a bottleneck even for large-scale models like InternVL3-78B (27.24%) and Qwen2.5-VL-72B (44.98%), suggesting persistent challenges in fine-grained spatial-semantic alignment.

Reasoning: High-Level Generalization. Reasoning tasks are the most demanding. Closed-source models such as GPT-4o and Gemini2.5-Pro achieve strong results (51.14%/56.20% on SRC and 55.20%/59.31% on SKI). Among open-source models, Qwen2.5-VL-72B leads, while the reasoning-specialized QVQ-Max approaches closed-source performance (58.86% on SKI) despite lacking OD and VG capabilities. This suggests that explicit reasoning models can partially compensate for perceptual limitations, likely relying on test-time scaling rather than grounded perception.

4.4 Synergistic effects analysis

To analyze the cross-task performance patterns of different models, we perform a statistical analysis of the synergies between different tasks and visualized the results as in Figure 6. We compute the Pearson correlation coefficients between Perception and Understanding tasks and observe notable interdependencies. OC and RE exhibit a strong positive correlation of 0.73, while OE and OCR show a similarly significant correlation of 0.67. These results indicate that effective performance in perception directly contributes to understanding,



(a) Perception, understanding, and reasoning correlation matrix

(b) Scatter plots between perception, understanding, and reasoning

(c) Linear regression analyses on reasoning tasks

(d) The feature importance analysis of regression coefficients

Figure 6: Statistical analysis of model accuracy and synergies between different tasks.

which in turn underpins downstream reasoning. Scatter plot visualizations further confirm these links, OCR, in particular, correlates strongly with both SRC and SKI, underscoring its central role in enabling semantic reasoning. Linear regression analyses reinforce these findings: OE and OCR are strong predictors of SRC, while OC and RE significantly influence SKI, highlighting how object-level detection and relational reasoning jointly support high-level inference. Finally, we apply second-order polynomial regression and the feature importance analysis of regression coefficients reveals task-specific contributions. These insights collectively demonstrate the layered structure of visual reasoning pipelines, where perception and understanding stages must be well-aligned to support robust inference. For further analysis, please refer to the appendix A.7.

4.5 Effectiveness of SMEC

To further evaluate the effectiveness of our proposed SMEC framework, we conduct experiments on the Scene Knowledge Inference (SKI) task using both a sampled subset of 3,500 question–answer pairs and the full test set (Table 3). Compared to the direct inference baseline, SMEC consistently improves performance across both model scales and different iteration depths. For Qwen2.5VL-7B, accuracy increases from 39.80% to 43.24% as the number of iterations N_t grows from 1 to 3, demonstrating the benefits of multi-step expert collaboration. Notably, SMEC also outperforms Self-Refine (Madaan et al., 2023) and Majority voting (Chen et al., 2024b) under the same setting. A similar trend appears at larger scales. With Qwen2.5VL-32B, SMEC improves accuracy from 49.17% to 52.44% with three iterations, confirming that the benefits of iterative expert collaboration scale with model capacity. Importantly, when evaluated on the full test set rather than the 3,500-sample subset, SMEC continues to provide consistent improvements (+3.12%). This indicates that SMEC's gains are not distribution-specific and remain stable under more comprehensive evaluation conditions.

Table 3: Accuracy comparison with different settings.

Methods	Model	Iterations	Performance
Direct	Qwen2.5VL-7b	-	39.80
Majority voting	Qwen2.5VL-7b	-	40.66
Self-Refine	Qwen2.5VL-7b	-	40.51 (+0.71)
SMEC	Qwen2.5VL-7b	1	41.35 (+1.55)
SMEC	Qwen2.5VL-7b	2	42.97 (+3.17)
SMEC	Qwen2.5VL-7b	3	43.24 (+3.44)
Direct	Qwen2.5VL-32b	-	49.17
SMEC	Qwen2.5VL-32b	3	52.44 (+3.27)
Direct (Full data)	Qwen2.5VL-32b	-	51.54
SMEC (Full data)	Qwen2.5VL-32b	3	54.66 (+3.12)

5 CONCLUSION

We contribute Lens, a multi-level benchmark designed to evaluate Multimodal Large Language Models (MLLMs) across perception, understanding, and reasoning. Unlike prior benchmarks, Lens aligns all tasks to the same set of realistic, contemporary images, enabling fine-grained analysis of how low-level visual capabilities support higher-order reasoning. The evaluation of recent MLLMs further reveals a consistent performance gap in reasoning tasks, highlighting the limitations of current models in integrating perception and cognition. To address this, we proposed SMEC, a self-driven multi-expert collaborative framework that prompts the MLLM to simulate a panel of specialized agents. Together, Lens and SMEC offer a new paradigm for evaluating and enhancing reasoning intelligence in MLLMs, paving the way to more robust, human-aligned multimodal intelligence.

6 ETHICS STATEMENT

Data Collection and Privacy. All images in the Lens dataset were collected manually from publicly available posts on social media platforms, including X (formerly Twitter), Instagram, Weibo, and RedNote. We strictly complied with the developer agreements and copyright/licensing regulations of these platforms. No private or restricted data was accessed, and all collection adhered to academic research policies.

Human Annotation. The dataset was constructed with the assistance of more than 50 undergraduate and graduate student annotators, who were trained to ensure annotation quality and consistency. All annotators were fairly compensated for their work. The annotation process involved only task-related labeling (*e.g.*, bounding boxes, question-answer generation) and did not involve collection of personal or sensitive information about the annotators.

Bias, Fairness, and Representation. Although images were sourced from global platforms to ensure diversity of cultural and regional content, dataset bias may still exist due to platform-level demographic imbalances and uneven scenario representation. We acknowledge that these limitations may influence model evaluation results, and we encourage future work to further broaden geographic and cultural coverage.

Research Integrity and Transparency. This study follows established standards of research integrity. We provide detailed dataset descriptions, evaluation protocols, and implementation details. We have no conflicts of interest or external sponsorship that might bias the study.

7 REPRODUCIBILITY STATEMENT

We have taken extensive measures to ensure the reproducibility of our results. A complete description of the Lens dataset, including data collection principles, filtering steps, and annotation procedures, is provided in Section 2 and Appendix. Details of the evaluation tasks, metrics, and benchmark comparisons are reported in Section 4, with additional analyses in Appendix. The implementation of our proposed Self-Driven Multi-Expert Collaborative Framework (SMEC), including the expert construction, screening, and integration process, is fully described in Section 3 and formalized in Appendix A.9. Hyperparameter settings, model configurations, and ablation experiments are included in the appendix to facilitate replication of results. Together, these resources ensure that both dataset construction and methodological contributions can be faithfully reproduced by the research community.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint*

arXiv:2303.08774, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*, 2023a.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*, 2024a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei A Zaharia, and James Y Zou. Are more llm calls all you need? towards the scaling properties of compound ai systems. *Advances in Neural Information Processing Systems*, 37:45767–45790, 2024b.

Wei Chen, Long Chen, and Yu Wu. An efficient and effective transformer decoder-based framework for multi-task visual grounding. In *European Conference on Computer Vision*, pp. 125–141, 2024c.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*, 2023c.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024.

Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *Advances in neural information processing systems*, 37:121670–121698, 2024.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1769–1779, 2021.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. *arXiv preprint arXiv:2504.01017*, 2025.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024.

Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Multi-modal agent tuning: Building a VLM-driven agent for efficient tool usage. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0bmGL4q7vJ.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14953–14962, 2023.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2023.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Yifan Jiang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, Jay Pujara, et al. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems*, 37:46567–46592, 2024.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*, 2024a.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024b.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14963–14973, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*, 2023a.

Fan Liu, Wenshuo Chao, Naiqiang Tan, and Hao Liu. Bag of tricks for inference-time computation of llm reasoning. *arXiv preprint arXiv:2502.07191*, 2025a.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023b.

Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. *arXiv preprint arXiv:2409.14750*, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024b.

Weiwen Liu, Xu Huang, Xingshan Zeng, xinlong hao, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Zhengying Liu, Yuanqing Yu, Zezhong WANG, Yuxian Wang, Wu Ning, Yutai Hou, Bin Wang, Chuhan Wu, Wang Xinzhi, Yong Liu, Yasheng Wang, Duyu Tang, Dandan Tu, Lifeng Shang, Xin Jiang, Ruiming Tang, Defu Lian, Qun Liu, and Enhong Chen. ToolACE: Winning the points of LLM function calling. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=8EB8k6DdCU.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024d.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025c.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pp. 417–435. Springer, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13279–13288, 2024.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Pavan Kartheek Rachabatuni, Filippo Principi, Paolo Mazzanti, and Marco Bertini. Context-aware chatbot using mllms for cultural heritage. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pp. 459–463, 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Gabriel Sarch, Lawrence Jang, Michael Tarr, William W Cohen, Kenneth Marino, and Katerina Fragkiadaki. Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *Advances in Neural Information Processing Systems*, 37:75942–75985, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025a.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6613–6629, 2025b.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025a.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025b.

Yunjie Tian, Tianren Ma, Lingxi Xie, and Qixiang Ye. Chatterbox: Multimodal referring and grounding with chain-of-questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7401–7409, Apr. 2025. doi: 10.1609/aaai.v39i7.32796. URL https://ojs.aaai.org/index.php/AAAI/article/view/32796.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Renxi Wang, Xudong Han, Lei Ji, Shu Wang, Timothy Baldwin, and Haonan Li. Toolgen: Unified tool retrieval and calling via generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XLMAMmowdY.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pp. 471–490. Springer, 2024a.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *European Conference on Computer Vision*, pp. 288–304. Springer, 2024b.

Fangyun Wei, Jinjing Zhao, Kun Yan, Hongyang Zhang, and Chang Xu. A large-scale human-centric benchmark for referring expression comprehension in the lmm era. *Advances in Neural Information Processing Systems*, 37:69566–69587, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 26:4334–4347, 2023.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *arXiv preprint arXiv:2412.20206*, 2024.

Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. *arXiv preprint arXiv:2410.12332*, 2024.

Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9499–9508, 2022.

Ruilin Yao, Shengwu Xiong, Yichen Zhao, and Yi Rong. Visual grounding with multi-modal conditional adaptation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3877–3886, 2024.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2msbbX3ydD.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85. Springer, 2016.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Kan Ren, Dongsheng Li, and Deqing Yang. Easytool: Enhancing llm-based agents with concise tool instruction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 951–972, 2025.

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

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 133(2):825–843, 2025.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025a.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture. *arXiv preprint arXiv:2509.02359*, 2025b.

Yikai Zhang, Siyu Yuan, Caiyu Hu, Kyle Richardson, Yanghua Xiao, and Jiangjie Chen. Timearena: Shaping efficient multitasking language agents in a time-aware simulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3894–3916, 2024a.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025c.

Zhehao Zhang, Ryan Rossi, Tong Yu, Franck Dernoncourt, Ruiyi Zhang, Jiuxiang Gu, Sungchul Kim, Xiang Chen, Zichao Wang, and Nedim Lipka. Vipact: Visual-perception enhancement via specialized vlm agent collaboration and tool-use. *arXiv preprint arXiv:2410.16400*, 2024b.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024c.

Chuyang Zhao, YuXin Song, Junru Chen, Kang Rong, Haocheng Feng, Gang Zhang, Shufan Ji, Jingdong Wang, Errui Ding, and Yifan Sun. Octopus: A multi-modal llm with parallel recognition and sequential understanding. *Advances in Neural Information Processing Systems*, 37:90009–90029, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *Advances in Neural Information Processing Systems*, 37:103305–103333, 2024.

A Appendix

A.1 Related work

A.1.1 Benchmarks for Visual Capability of MLLMs

The capability of Visual Perception, Understanding and Reasoning is a foundational aspect of understanding benchmarks, which involves the ability to recognize and localize multiple objects, interpret various visual elements with complex emotional or implicit cues and summarize visual information for feedback and decision making (Li et al., 2024b). Specifically, Perception in MLLMs involves the classification, detection of basic visual objects (*e.g.*, dog, cat) and attributes (*e.g.*, color, lighting). These low-level perceptual capabilities are crucial for various applications, including recognition systems (Zhao et al., 2024) and visual quality enhancement (Zhang et al., 2024c). Understanding represents a sophisticated level of image understanding that focuses on the detailed and nuanced aspects of visual content. It includes recognizing and interpreting the visual-linguistic concepts, such as text recognition (OCRBench (Liu et al., 2024d)), Visual Grounding (RefCOCO (Yu et al., 2016), FineCops-Ref (Liu et al., 2024a), HC-RefLoCo (Wei et al., 2024)) and Referring Expression Generation (Visual Genome) (Krishna et al., 2017), which refers to the model's ability to accurately link visual elements with corresponding textual descriptions. Although tasks at this level begin to involve visual and textual alignment, they still do not require reasoning or external knowledge. For higher-order capability, reasoning in MLLMs involves advanced event understanding and deep meaning extraction from multimodal data. These capabilities include interpreting and responding to complex emotional cues across multiple modalities (Cheng et al., 2024), deriving subtle implicit meanings from visual and contextual information (Liu et al., 2023a), and a range of other competencies, including knowledge acquisition, language generation, spatial awareness, and cultural context integration (Rachabatuni et al., 2024).

A.1.2 Reasoning Capability of MLLMs

MLLMs have demonstrated remarkable reasoning capabilities, largely facilitated by test-time scaling (Dong et al., 2022; Wei et al., 2022), which allows feeding prompted samples and context. This capability has been further enhanced by chain-of-thought (CoT) prompting (Wei et al., 2022), which enables LLMs to generate coherent intermediate reasoning steps toward the final answer. Previous studies have shown that LLMs benefit from manually written demonstrations as well as zero-shot prompting outputs. However, due to the domain gap between various modalities, the current reasoning capability of MLLMs in the complex real-world environment is still limited. To address this limitation, researchers have focused on enhancing the reasoning capability of MLLMs in both the training and prompting paradigms. Flamingo (Alayrac et al., 2022) bridges the gap between these two modalities by pre-training on interleaved visual and textual data. Some other works, such as Shikra (Chen et al., 2023b) and Ferret (You et al., 2023), leverage visual grounding data (Xiao et al., 2024; Yao et al., 2024) to achieve fine-grained vision-language alignment. Furthermore, recent studies have also demonstrated that augmenting computing resources during the testing phase (test-time scaling) can enhance the reasoning capabilities of LLMs (Jaech et al., 2024). More specifically, Prompt-based Reasoning Meta-Systems (PRMS) can be employed to guide LLMs in evaluating and filtering intermediate "thinking" processes (Snell et al., 2024). This encourages the generation of more sophisticated reasoning steps during testing, ultimately leading to improved reasoning accuracy. Beyond that, some methods employ the external knowledge to focus on important visual details, like V* (Wu & Xie, 2024), Marvel (Jiang et al., 2024), and ICAL (Sarch et al., 2024), collecting a series of visual reasoning steps as training data. More recently, with the emergence of DeepSeek-R1 (Guo et al., 2025) demonstrating strong potential in LLM reasoning, research efforts have begun to explore reasoning-centric models and R1-style reinforcement learning strategies for understanding complex visual scenes and tasks. These studies (Huang et al., 2025; Shen et al., 2025a; Liu et al., 2025c) particularly emphasize the long-chain reasoning capabilities within MLLMs, aiming to enhance their performance in handling intricate visual-linguistic reasoning challenges.

A.1.3 LLM-BASED AGENTIC REASONING

The rapid progress of large language models (Achiam et al., 2023; Bai et al., 2025) has sparked significant interest in building autonomous agents capable of solving complex, multi-step reasoning tasks. Leveraging the strong chain-of-thought (CoT) abilities of modern LLMs (Wei et al., 2022), these systems typically decompose a complex problem into a sequence of structured subtasks, invoke intermediate deliberation, and integrate the resulting insights to produce a final answer (Gupta & Kembhavi, 2023; Chen et al., 2023a). Recent developments in LLM-based autonomous agents highlight the importance of planning (Huang et al., 2024; Zhang et al., 2024a), tool usage (Yuan et al., 2025), memory (Zhang et al., 2025c), and persona (Chen et al., 2024a). In parallel, multi-agent frameworks such as MetaGPT (Hong et al., 2023), AgentVerse (Chen et al., 2023c), etc., demonstrate strong performance by orchestrating multiple interacting agents, often instantiated as distinct roles with specialized responsibilities. Despite their success, these systems rely heavily on manually designed personas, fixed role hierarchies, or hand-crafted coordination rules, which limits their flexibility and generalization across tasks and domains. Moreover, the dependence on external scaffolding or pre-specified agent behaviors often restricts the model's ability to adaptively adjust its internal reasoning pathway. To address these limitations, SMEC introduces a self-driven agentic reasoning mechanism that automatically generates diverse experts, elicits their reasoning, filters redundant experts, and synthesizes their perspectives into a final consensus, all within the model's own language-native inference loop.

A.2 SPECIFIC DEFINATION OF DIFFERENT TASKS

Object Counting (OC): Estimating the number of object instances described by a free-form expression, often under complex conditions like occlusion, scale variation, or clutter.

Object Detection (OD): Localizing objects within an image by generating bounding boxes paired with corresponding class labels. In order to better match the real-life scenarios and practical applications, we construct more than 500 fine-grained object categories based on natural language.

Object Existence Determination (OE): Determining whether a particular object, which described by a detailed expression, exists in the image without requiring spatial localization.

Relation Extraction (RE): Identifying semantic relationships (*e.g.*, "holding", "next to", "wearing") between pairs of objects to facilitate structured scene understanding. And we added questions about the objects that do not exist in the images to evaluate model's ability to suppress hallucinations.

Visual Grounding (VG): Localizing an image region that corresponds to a natural language expression, linking linguistic references to fine-grained visual content.

Region-wise OCR (OCR): Recognizing and transcribing text within a region, which specified by coordinates or description, facilitating fine-grained interleaved image-text understanding.

Spatial Relationship Comprehension (SRC): Understanding geometric relationships (*e.g.*, "above" and "to the left front to") between objects within diverse 3D views, supporting visual-spatial reasoning. Compared to some rudimentary or synthetic spatial understanding datasets (Johnson et al., 2017; Li et al., 2023; Liu et al., 2023b), our data is more realistic in emphasizing spatial location understanding under real-world scenarios as well as 2D images acquired by cameras or cell phones.

Scene Knowledge Inference (SKI): Inferring high-level semantic and functional information about the scene or making decision based on the visual contents, incorporating context, commonsense knowledge, and visual cues beyond explicit visual entities. Compared to the regular visual reasoning dataset, `Lens` additionally distinguish between "thought paths" and "final answers", differentiated by the $< think >$ token, aiming to provide richer and finer-grained information for potential test-time scaling tests and R1-style reinforcement learning.

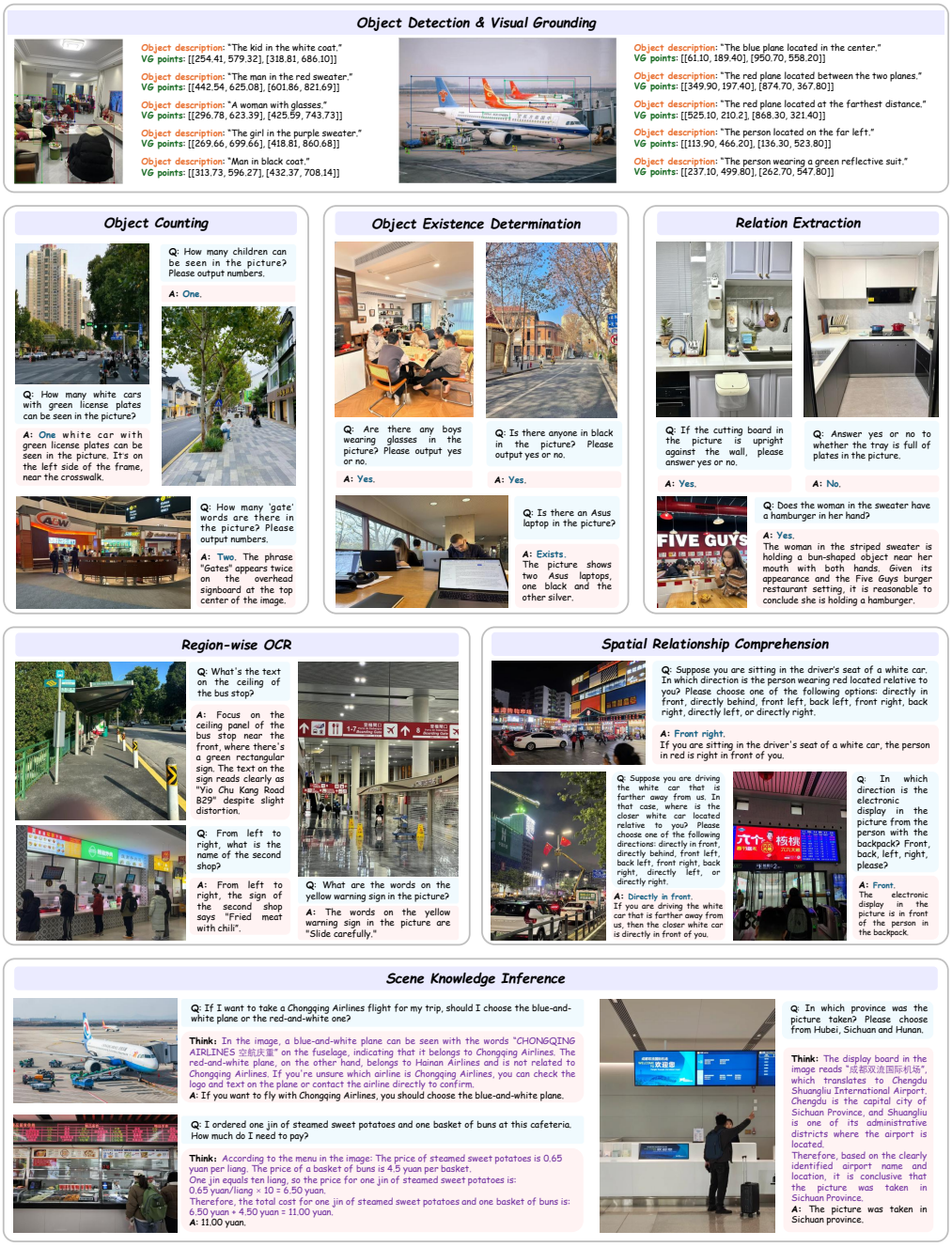


Figure 7: Lens covers a wide range of images and annotations, from fine-grained recognition and spatial localization to complex reasoning over extended thought processes. Notably, each image is annotated with labels corresponding to all subtasks concurrently, enabling comprehensive evaluation.

A.3 Quality control process

In addition to the annotations, diversified measures were taken to enrich the content of data samples and ensure their quality. Specifically, we implemented a multi-faceted quality control process. Beyond a two-step data cleaning protocol, we also enriched each image with supplementary metadata to facilitate traceability and contextual analysis.

First, we manually performed a two-stage data cleaning process. In the initial stage, we reviewed and eliminated suspected duplicate images. The second stage involved distributing the problems among co-authors for meticulous format and typo checking, ensuring all annotations adhered to a standardized format.

To further validate the quality and consistency of our annotations, we performed an additional two-step verification process. This included both manual and machine-assisted checks. The entire dataset was cross-verified by both an independent team of annotators and the Qwen2.5-VL 72B open-source model. For machine validation, we input the original image, question, and answer into the MLLM. Cases flagged as invalid by the model were isolated for manual re-evaluation by a separate team of annotators. For object detection and visual grounding tasks, we directly visualized the annotations on the images, enabling human evaluators to assess the validity of the bounding boxes.

Additionally, we enriched each image with supplementary metadata. We included a pseudonymized Annotator ID to allow for annotator-specific quality tracking while preserving privacy. The Time of Online Publication and Scene Category were also labeled to facilitate temporal studies, filter outdated content, and organize the dataset by scene. Finally, ambiguous images that consistently resulted in low inter-annotator agreement were manually filtered out to ensure a high-quality final dataset. Some cleaned representative examples are visualized in Figure 7.

A.4 Data Privacy Protection and Copyright Statement

Our protocol for handling potentially sensitive information was a multi-stage process designed to be as thorough as possible:

Automated Pre-screening: As an initial step, we used automated tools (*e.g.*, face detection models [3] , docTR [4]) to perform a preliminary scan of the collected images. This scan was configured to flag images with a high probability of containing human faces or dense blocks of text that might constitute personally identifiable information.

Comprehensive Manual Review: Every image, including those not flagged by the automated scan, was then subjected to a thorough manual review by our team of over 20 trained human annotators. Annotators received specific training and a detailed guide on identifying a wide range of sensitive data, including but not limited to: Visible and recognizable faces; Full names, usernames, or contact information; License plates, street addresses, or other specific location markers; Private documents or screens displaying personal data.

Sensitive Information Exclusion: Based on the manual review, if an image contained sensitive information, one of two actions was taken as mentioned in the paper:

- Processing: If the sensitive information was incidental to the image's main content, we applied irreversible blurring or masking to the specific region.
- Exclusion: If the sensitive information was central to the image and could not be adequately anonymized without destroying the scene's context, the image was entirely excluded from the final dataset.

[3] https://github.com/timesler/facenet-pytorch
[4] https://github.com/mindee/doctr

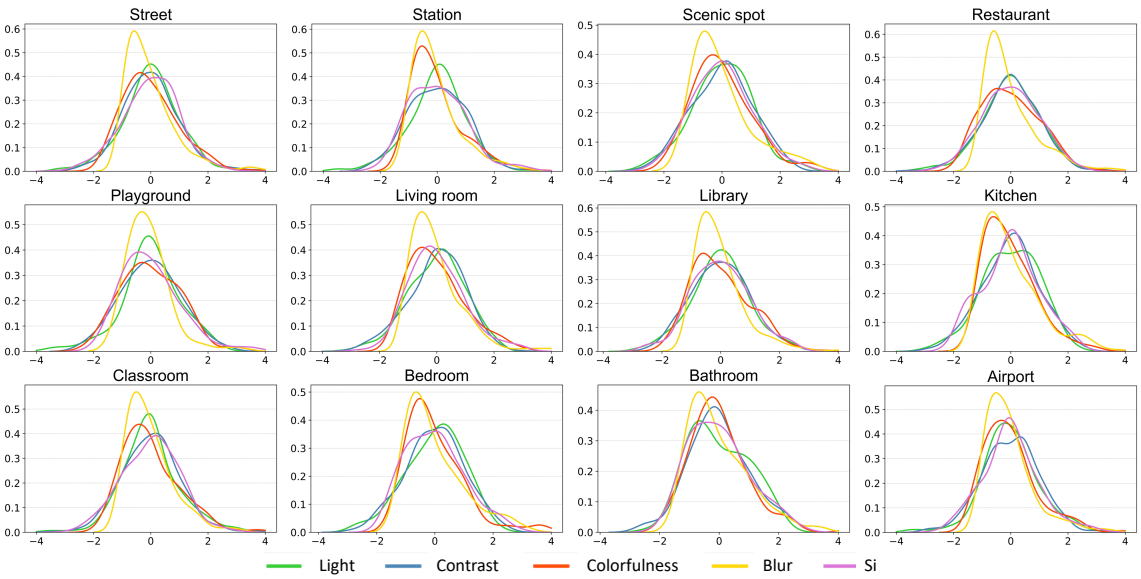


Figure 8: The normalized probability distributions of low-level attributes from different scenes. Scenes with flat peaks show more diversity, while those with sharp peaks have similar features.

Final Verification: To ensure consistency and quality, a final audit was conducted by a subset of the paper's authors. This team reviewed a random sample of the approved images and 100% of the processed (blurred/masked) images to verify that our privacy protocol was correctly and consistently applied.

This multi-stage, human-centric approach ensures that the images in the `Lens` dataset comply with platform policies and respect individual privacy.

A.5 Low-level feature analysis of images from different scenes

We counted five low-level visual attributes, including lighting, contrast, color, blur, and spatial information (SI), to assess the statistical difference between different scenes. As shown in Figure 8, the normalized probability density curves of low-level visual attributes across different scenes are consistent with human perceptual preferences. Scenes with regulated lighting conditions (*e.g.*, classrooms, airports, and stations) demonstrate sharp peaks near x $\approx$ 0 in the illumination curves (density $>$ 0.5), indicating constrained variations in brightness. In contrast, domestic environments (*e.g.*, living rooms, bedrooms, and kitchens) display broader illumination distributions, suggesting more diverse and adaptive light sources. Furthermore, functional scenes such as bedrooms, bathrooms, and kitchens exhibit sharp, concentrated peaks in color distributions (peak density $\approx$ 0.5), implying greater structural regularity in specific visual attributes.

A.6 Fine-grained evaluation of VG

We conducted a fine-grained evaluation of a diverse set of models on the Visual Grounding (VG) task, categorizing them into two groups: traditional predictive multimodal models and generative multimodal large language models (MLLMs). The results, summarized in Table 4, reveal several key insights into the current state of visual grounding capabilities.

A.6.1 Comparison of Model Categories

The results clearly indicate a significant performance gap between the two model categories. The top-performing MLLMs, specifically Qwen2.5-VL-7B and Qwen2.5-VL-32B, demonstrate superior performance across all metrics, with an accuracy of 46.94% and 48.47% at $IoU@0.5$, respectively. This performance is substantially higher than the top predictive model, G-DINO, which achieves 37.05% at the same metric. This finding suggests that the generative and in-context learning capabilities of modern MLLMs provide a substantial advantage in the complex VG task, enabling them to better understand nuanced linguistic instructions and ground them accurately in the visual space.

A.6.2 Performance on Different Scales

A multi-scale analysis, measured by accuracy on small (ACC_s), medium (ACC_m), and large (ACC_l) objects, provides a more granular view of each model's strengths and weaknesses. Both traditional and generative models exhibit a similar trend: performance consistently improves with the size of the target object. This is a common challenge in visual grounding and object detection, as localizing and grounding small objects remains difficult.

Traditional Models: Among the traditional models, G-DINO demonstrates a more balanced performance across scales, achieving 24.87% on small objects and 52.32% on large objects. In contrast, models like VLTVG and SimVG struggle significantly with small objects, with accuracies of 0.00% and 0.01% respectively, but show strong performance on large objects (29.70% and 45.20%).

Generative MLLMs: While MLLMs also struggle with small objects, their performance is notably better than most traditional models. Qwen2.5-VL-32B and Qwen2.5-VL-7B achieve high accuracy on medium and large objects, with their ACC_m and ACC_l scores reaching 55.04% and 61.48% (for Qwen2.5-VL-32B), and 54.12% and 60.24% (for Qwen2.5-VL-7B) respectively. The strong performance on larger objects may be attributed to their powerful visual backbones and advanced language understanding capabilities, which help them better contextualize the target within the scene.

A.6.3 Impact of Model Architecture and Size

Our results also highlight the importance of model architecture and size. The Qwen2.5-VL family of models, with its impressive performance, benefits from a powerful visual encoder (FE-ViT) and a sophisticated Qwen2.5 language backbone. Similarly, the InternVL3 series shows a clear scaling effect, where performance on most metrics improves as the model size increases from 2B to 14B. The performance of the 38B variant is slightly lower than the 14B variant due to its different visual backbone. This trend, consistent with findings in large language models, suggests that scaling up both visual and linguistic components is a promising direction for future research in visual grounding.

A.7 More Synergistic Effects Analysis

We further notice that Figure 6 (a) shows OC-OCR correlation (0.77) $\gg$ OC-OE (0.46). This contradicts intuition—object counting (OC) should align more naturally with existence checks (OE) than OCR. We attribute this to two factors. First, the Object Existence (OE) task is a simple binary classification: either an object is present or it is not. It requires a model to make a broad, scene-level assessment. In contrast, Object Counting (OC) is a more demanding task that requires fine-grained localization of individual objects, followed by an enumeration step. A model can be highly proficient at a binary existence check without possessing the precise localization and counting skills needed for the OC task. This fundamental difference in cognitive demand limits the correlation between the two. Second, the high correlation between Object Counting (OC) and OCR is not coincidental. Both tasks rely on a critical shared capability: fine-grained localization. To perform well on the OC task, a model must accurately identify and localize each instance of an object to

Method	Visual Backbone	Linguistic Backbone	Accuracy @ IoU @0.5	@0.6	@0.7	@0.8	@0.9	Scale-wise Accuracy ACC_s	ACC_m	ACC_l
Methods based on predictive multimodal models:										
TransVG (Deng et al., 2021)	RN101	BERT-B	8.73	7.57	6.29	4.40	1.69	0.01	2.01	23.64
VLTVG (Yang et al., 2022)	RN101	BERT-B	11.04	9.60	7.75	5.33	1.99	0.00	2.80	29.70
MMCA (Yao et al., 2024)	RN101	BERT-B	10.92	9.45	7.90	5.64	2.22	0.03	2.79	29.31
CLIP-VG (Xiao et al., 2023)	CLIP-B	CLIP-B	8.73	7.57	6.29	4.40	1.69	0.01	2.01	23.64
EEVG (Chen et al., 2024c)	ViT-B/16	BERT-B	9.27	5.78	2.51	0.48	0.05	0.01	0.98	26.12
SimVG (Dai et al., 2024)	BEIT-3	BEIT-3	16.46	13.90	11.12	7.44	2.70	0.01	3.10	45.20
G-DINO (Liu et al., 2024b)	Swin-L	BERT-B	37.05	33.92	29.20	22.57	11.36	24.87	37.54	52.32
Methods based on generative multimodal large language models:										
Groma-7B (Ma et al., 2024)	DINOv2-L	Vicuna	33.59	29.95	25.47	18.73	8.52	11.58	33.91	58.59
Mova-7B (Zong et al., 2024)	Multi-expert	Vicuna	20.44	13.10	5.98	1.09	0.13	5.06	15.97	40.36
Ferret-7B (You et al., 2024)	CLIP-L	Vicuna	23.26	18.97	13.95	7.61	1.84	1.85	19.49	54.64
Ferret-13B (You et al., 2024)	CLIP-L	Vicuna	24.20	19.81	14.41	8.12	2.05	2.26	20.42	56.31
InternVL3-2B (Zhu et al., 2025)	InternViT-0.3B	Qwen2.5	7.89	5.10	2.85	1.36	0.33	0.61	3.46	19.34
InternVL3-8B (Zhu et al., 2025)	InternViT-0.3B	Qwen2.5	17.54	13.23	8.89	4.94	1.60	3.23	15.36	35.21
InternVL3-14B (Zhu et al., 2025)	InternViT-0.3B	Qwen2.5	29.53	23.98	17.25	10.05	3.00	4.58	27.80	57.07
InternVL3-38B (Zhu et al., 2025)	InternViT-6B	Qwen2.5	27.85	21.42	15.00	8.23	2.37	4.91	25.81	53.56
VLM-R1-3B (Shen et al., 2025a)	FE-ViT	Qwen2.5	23.79	19.91	15.65	10.53	4.31	8.15	22.84	40.94
Qwen2.5-VL-3B (Bai et al., 2025)	FE-ViT	Qwen2.5	45.03	37.92	29.33	18.48	6.51	29.14	50.54	57.15
Qwen2.5-VL-7B (Bai et al., 2025)	FE-ViT	Qwen2.5	46.94	39.39	29.94	18.38	6.26	28.87	54.12	60.24
Qwen2.5-VL-32B (Bai et al., 2025)	FE-ViT	Qwen2.5	48.47	40.66	30.78	19.15	6.63	30.93	55.04	61.48

Table 4: Multi-scale evaluation results.

count it. Similarly, to perform region-wise OCR, the model must first precisely locate the bounding box of the text before reading it. The strong correlation suggests that the ability to perform precise object localization is a dominant factor in a model's success on both tasks, thus strengthening their relationship despite their different end goals.

To further test the synergy between different tasks, we conducted a experiment with Qwen2.5-VL-7B on a sampled subset of `Lens`. Specifically, when testing the SKI task, we fed the VQA question-answer pairs of other tasks into the model as context along with the question, and asked it to return the answer (refer to Appendix A.15 for the prompt template p_s, where we provide an example based on the OCR task). The test results are shown in Table 5 and reveal several noteworthy patterns. First, incorporating OCR information yields a substantial performance gain (from 39.80% to 41.36%), indicating that understanding scene text helps the model solve some reasoning tasks. Second, although some tasks—such as OE and OC—exhibit limited or even negative effects when introduced individually, their combination with OCR consistently boosts performance. This may indicate that auxiliary perceptual signals, while insufficient on their own, can enhance reasoning when mediated through textual understanding. Furthermore, the observed synergistic effects resonate with the design philosophy of our proposed Self-Driven Multi-Expert Collaborative (SMEC) framework. We argue that complex multimodal reasoning cannot be achieved by isolated competencies alone, but requires a structured mechanism to coordinate heterogeneous sources of evidence.

OCR	-	✔	-	-	-	✔	✔	✔	✔	✔	✔
RE	-	-	✔	-	-	✔	-	-	✔	✔	✔
OE	-	-	-	✔	-	-	✔	-	✔	-	✔
OC	-	-	-	-	✔	-	-	✔	-	✔	✔
Performance	39.80	41.36	38.03	39.23	39.52	40.72	41.45	40.73	41.93	40.79	41.90

Table 5: Testing the synergistic effects of different tasks on Scene Knowledge Inference (SKI).

A.8 Analysis of Input Resolution

Method	Visual Backbone	Linguistic Backbone	Accuracy @ Input Resolution			
			640 × 640	960 × 960	1280 × 1280	1600 × 1600
InternVL3-2B (Zhu et al., 2025)	InternViT-0.3B	Qwen2.5	40.97	41.53	41.90	40.60
InternVL3-9B (Zhu et al., 2025)	InternViT-0.3B	InternLM3-8B	46.95	46.54	46.65	46.63
InternVL3-14B (Zhu et al., 2025)	InternViT-0.3B	Qwen2.5	50.28	51.15	51.58	51.17
InternVL3-38B (Zhu et al., 2025)	InternViT-6B	Qwen2.5	50.88	50.97	49.98	51.08
Qwen2.5-VL-3B (Bai et al., 2025)	FE-ViT	Qwen2.5	40.08	40.44	40.48	40.58
Qwen2.5-VL-7B (Bai et al., 2025)	FE-ViT	Qwen2.5	46.52	47.41	48.13	48.11
Qwen2.5-VL-32B (Bai et al., 2025)	FE-ViT	Qwen2.5	53.72	54.37	54.10	54.09
GLM-4.1V-Base-9B (Hong et al., 2025)	AIMv2-Huge	GLM-4-0414	42.35	42.86	43.28	43.73
GLM-4.1V-Thinking-9B (Hong et al., 2025)	AIMv2-Huge	GLM-4-0414	48.77	50.78	51.32	51.13

Table 6: Benchmark results across varying input resolutions.

We conducted a detailed analysis to understand the impact of varying input resolutions on model performance. The results, summarized in Table 6, reveal several key insights.

General Trend (Performance Improves with Resolution): For most models, performance generally improves as the input resolution increases. This trend is evident in models such as Qwen2.5-VL-7B, which shows a steady increase in accuracy from 46.52% at 640×640 to 48.13% at 1280x1280. Similarly, GLM-4.1V-Thinking-9B improves from 48.77% to 51.32% over the same range. This is expected, as higher resolutions provide more visual detail, which is particularly beneficial for complex visual grounding and reasoning tasks that require fine-grained perception.

The Point of Diminishing Returns: However, the results also suggest a point of diminishing returns. For many models, the performance gain from increasing the resolution beyond 1280x1280 is minimal, and in some cases, performance slightly decreases. For example, InternVL3-14B peaks at 51.58% at 1280x1280 and then slightly drops to 51.17% at 1600x1600. Similarly, Qwen2.5-VL-7B's performance plateaus at 1280×1280. This phenomenon could be attributed to several factors, including the model's architecture, which may not be fully optimized to handle the extra high-resolution information, or the fact that the added detail does not contribute meaningfully to solving the task.

Model-Specific Variations: Interestingly, some models, like InternVL3-2B, show less sensitivity to resolution changes, with its performance remaining relatively stable across all resolutions. In contrast, models such as GLM-4.1V-Thinking-9B and Qwen2.5-VL-32B demonstrate a more pronounced performance improvement with higher resolutions, indicating that their architectures are more capable of leveraging the extra visual information. This suggests that the optimal input resolution is not a one-size-fits-all solution and depends heavily on the model's architecture and design.

A.9 Qualitative error analysis

To better illustrate common failure patterns and the underlying limitations of current Multimodal Large Language Models, we conduct a qualitative analysis of representative error cases across different task levels. Following the structure of our benchmark, we group the visualizations into two categories: (1) VQA-style tasks and (2) Localization tasks, including detection and visual grounding, as shown in Figure 9.

For Perception & Understanding Tasks, they primarily require directly aligning visual content with textual queries. While modern MLLMs achieve reasonably high accuracy, their errors frequently stem from low-level perceptual limitations such as small objects, occlusion, distant subjects, and sensitivity to resolution. Several examples clearly illustrate this issue—for instance, miscounting the number of cyclists in a street scene, or failing to detect objects like a straw on a desk or a road light partially obscured by signage. For more complex reasoning tasks, we observe two major classes of systematic errors.

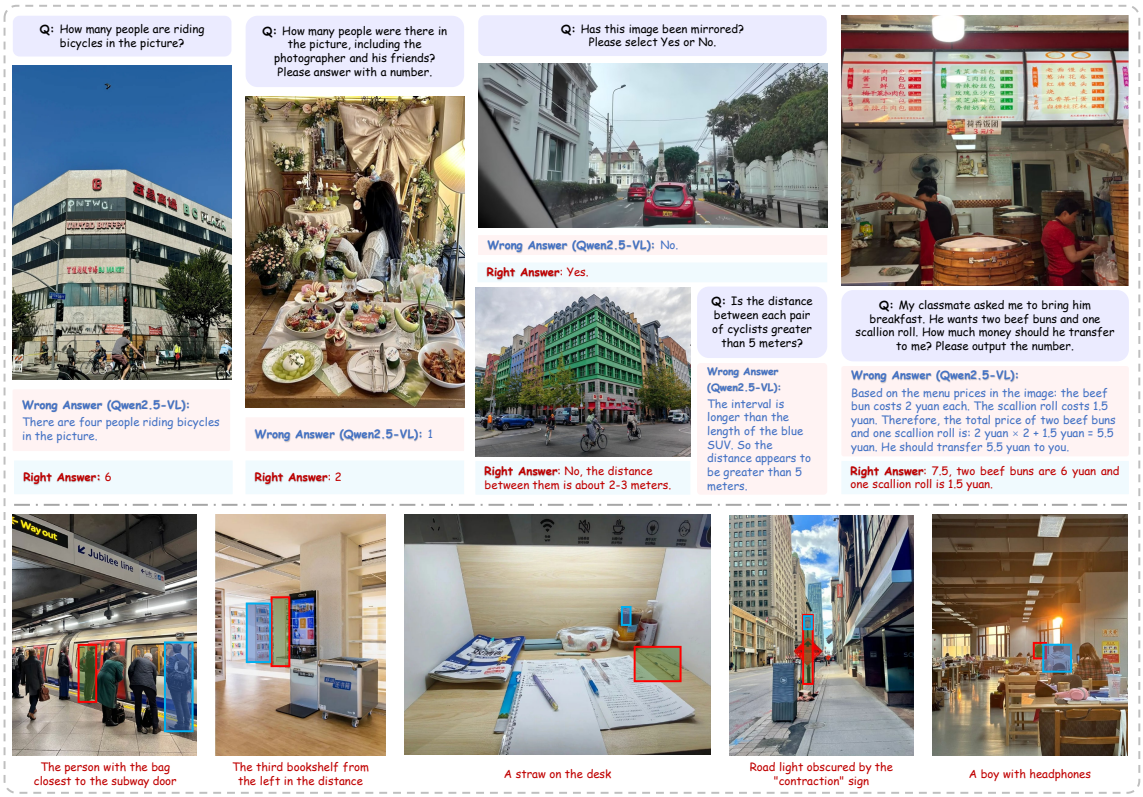


Figure 9: Failure cases for various tasks. Incorrect predictions and labels are indicated by blue and red, respectively

Correct reasoning, incorrect perception: The model often demonstrates sound logical reasoning but bases its inference on incorrect or incomplete visual extraction. For example, in a price-computation task, although the model performs the arithmetic correctly, it misreads the menu price of the beef bun, leading to a wrong total (5.5 instead of 7.5 yuan). This reveals a persistent bottleneck where high-level reasoning is constrained by low-level perception, especially OCR and fine-grained attribute recognition. Such cases motivate the need for agent-based or expert-collaborative pipelines—like our SMEC framework—which can iteratively refine visual cues or invoke specialized perception experts. They also highlight the importance of dynamic zoom-in strategies for capturing critical but small textual or visual details. (Wu & Xie, 2024; Zhang et al., 2025a; Shen et al., 2025b; Shao et al., 2024)

Spatial and physical reasoning deficits: A second recurring failure mode involves questions requiring geometric or physical commonsense. Models frequently struggle with tasks that implicitly require depth understanding, object-scale priors, or spatial metric reasoning. For instance, the model incorrectly concludes that two cyclists are more than 5 meters apart by comparing them to the length of a blue SUV, even though the correct distance is only about 2–3 meters. Likewise, it fails to judge whether an image is mirrored due to misunderstanding spatial layout cues. These issues echo recent findings showing that current MLLMs still lack robust spatial grounding, metric reasoning, and physical commonsense. (Zhang et al., 2025b; Azzolini et al., 2025)

A.10 A FORMAL DESCRIPTION OF SMEC

As shown in Algorithm 1. A key advantage of SMEC is that it does not depend on fixed, hand-crafted prompts. Instead, the prompts and expert descriptions are self-generated by the model based on the given visual input and question. During each iteration, the model adaptively refines its expert descriptions and updates the generation process when redundancy or low-quality information is detected. This adaptive design means that SMEC is not tied to a specific phrasing or a predefined set of experts, but can flexibly adjust to different problems and question types. As a result, our method is more robust than approaches that rely heavily on manually designed prompts, since the "experts" in SMEC emerge dynamically from the model itself rather than being externally imposed.

Algorithm 1 Self-driven Multi-Expert Generation & Collaboration

Initialization: Based Instruction-tuned MLLM θ, Question q, Meta Generation Prompt p_g, Inspection prompt p_i, Collaboration Prompt p_c, Description Set $D = \emptyset$, Maximum Answer Set $A = \emptyset$, Iterations N_t.

$a_0 = \theta(q), A = A \cup a_0$ # Initial answer for question.
for $t = 1, 2, \ldots, N_t$ **do**
  if $t = 1$ **then**
    $d_q^1 = \theta(p_g, q, A_0)$ # Initial expert description.
  else
    $d_q^t = \theta(p_g, D, q, A_t)$ # New description based on existing information.
  end if
  if $\theta(p_i, D, d_q^t) = Retain$ **then**
    $D = D \cup d_q^t$ # Checking process.
    $a_t = \theta(q, d_q^t), A = A \cup a_t$ # New answer from the expert perspective.
  else
    $p_g = \theta(q, d_q^t, p_g, D)$ # Update generation prompt while repeat descriptions.
  end if
end for
$a_{final} = \theta(q, A, p_c, D)$ # Summarize the final answer.

A.11 HUMAN PREFERENCE

To ensure the verifiability of our evaluation, particularly for open-ended reasoning tasks, we employed a large language model (LLM) as an automatic grader. To mitigate the concern regarding potential LLM hallucinations or failure to detect nuanced mistakes, it is noted that the LLM grader (*e.g.*, GLM4-flash) is used to compare the model-generated responses against our pre-existing, human-annotated answers, ensuring that the ground truth remains anchored in high-quality human data. We also conducted a human preference analysis on a representative and complex task SKI with a subset of the dataset, aiming to provide a gold standard against which to measure the reliability of judgement model. We recruited a separate team of ten expert annotators who were not involved in the original data collection to manually evaluate the accuracy of the model's answers compared to the labeled answers, as shown in Table 7. In this setup, Actual Positive (AP) and Actual Negative (AN) represent human judgments of correctness and incorrectness, respectively, while Test Positive and Test Negative indicate the LLM grader's corresponding evaluations.

The results, shown in Table 7, demonstrate a strong alignment between human preference and LLM-based judgments across all evaluated models. For instance, in the case of Qwen2.5-VL's responses, when humans

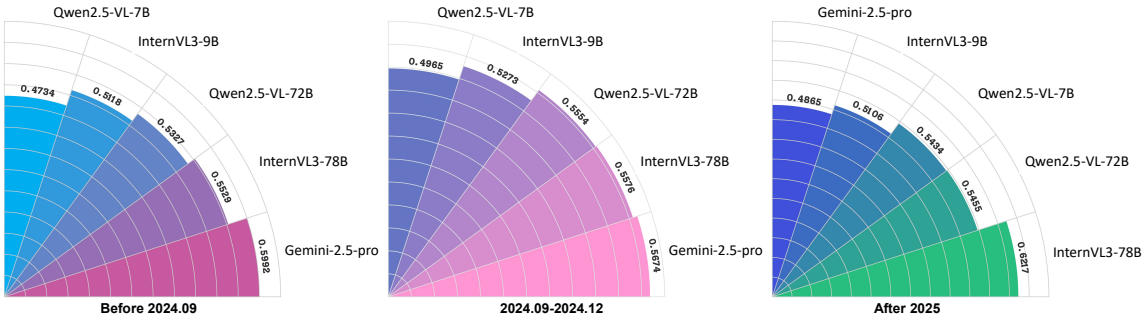


Figure 10: Model accuracy on SKI task across temporal split.

labeled an answer as correct (AP), the LLM grader also marked it as correct 97.13% of the time. Similarly, when humans judged an answer as incorrect (AN), the LLM grader agreed 96.14% of the time. Comparable trends are observed for InternVL3 and Gemini2.5-pro, though with slightly larger gaps in negative cases. These findings suggest that the LLM grader provides a highly reliable approximation of human judgment, especially for positive cases. Incorporating human validation thus not only confirms the feasibility of using LLMs as evaluators but also highlights their potential to scale evaluation consistently across large datasets while retaining strong alignment with expert human preference.

Method	**Qwen2.5-VL**		**InternVL3**		**Gemini2.5-pro**	
	AP	**AN**	**AP**	**AN**	**AP**	**AN**
Test Positive	97.13%	2.87%	94.70%	5.30%	93.65%	6.35%
Test Negative	3.86%	96.14%	7.77%	92.23%	11.41%	88.59%

Table 7: Human evaluation for the models' responses.

A.12 Temporal Generalization Performance

We visualize the accuracy of the best few models on data from different time periods in Figure 10, including Qwen2.5-VL (7B/72B), InternVL3 (9B/78B), and Gemini-2.5-Pro. Notably, models released after late 2024, such as InternVL3-78B and Qwen2.5-VL-72B, consistently outperform Gemini-2.5-pro on contemporary data, although their overall performance lower than Gemini-2.5-pro. The experimental results further supporting the observation that newer models tend to generalize better to new data distributions. This trend underscores the impact of scaling, instruction tuning, and exposure to temporally aligned data in enhancing multimodal reasoning performance.

A.13 More on SMEC vs. Baseline methods

To further highlight the advantage of our proposed SMEC framework, we provide additional qualitative comparisons against baseline methods, including direct inference and ChatGPT-style single-pass reasoning. Figure 11 showcases representative challenging samples drawn from `Lens`, where baseline models tend to produce either incomplete or overconfident predictions.

In these cases, ChatGPT and other baselines often failed for two recurring reasons:

Over-Reliance on Surface Cues. Baselines typically produced answers anchored on the most salient visual elements, neglecting contextual or relational signals. For instance, when asked to infer spatial constraints or traffic rules, ChatGPT tended to extrapolate directly from textual OCR cues, leading to plausible but incorrect answers.

Lack of Internal Deliberation. Without multi-perspective reasoning, baselines converged prematurely on a single hypothesis. This often caused brittle errors in scenarios requiring integration of textual, spatial, and commonsense evidence.

By contrast, SMEC decomposed the problem into multiple role-specific perspectives, such as a scene analyst, a spatial reasoner, and a cultural or commonsense expert. These experts generated partially overlapping but complementary hypotheses, which were then screened for redundancy and synthesized into a consensus. In the traffic-sign example shown in Figure 11, SMEC correctly filtered out spurious cues and converged on the right driving instruction, whereas ChatGPT remained uncertain or hallucinated unsupported details. Across difficult samples, two consistent patterns emerged:

Error Correction through Redundancy Filtering. Even when some experts produced misleading interpretations, SMEC's screening stage effectively down-weighted unreliable reasoning chains, preventing error propagation.

Multi-Step Enrichment. Iterative expert collaboration yielded richer reasoning trajectories, allowing the framework to exploit synergies between visual grounding, OCR, and commonsense inference. This process systematically improved robustness to ambiguous or noisy inputs.

Taken together, these results suggest that SMEC is not merely a test-time ensemble but a principled framework that encourages internal debate within a single model. Unlike single-pass inference, SMEC operationalizes a language-native form of deliberation, enabling models to approximate the dynamics of human expert panels. As Figure 11 illustrates, this mechanism directly translates into more accurate and interpretable reasoning on complex multimodal questions.

A.14 LIMITATIONS

While `Lens` offers broad task coverage and a unified evaluation setting, it currently focuses on static images and short-form reasoning. Real-world applications may require multimodal reasoning over temporal sequences or long-form narratives, which are beyond the scope of this version. Additionally, SMEC relies on prompt-based expert simulation, which, though flexible, may introduce redundancy or sensitivity to prompt phrasing.

A.15 PROMPTS FOR SYNERGY TESTING

Prompt p_s with OCR question-answer pairs as context

$\{question\ from\ SKI\}$
You can use the following facts to help you answer this question. Please note that they may not be relevant to the question. Here are some factual questions and answers about this picture:
$\{question\ from\ OCR\}$: $\{answer\ from\ OCR\}$

Q: Is my floor the 2nd floor?

Direct Answer (from Chatgpt):

From the signs in your photo, you are currently on Level 2.
I can see the directory on the left which shows:
Level 5: Departure Hall
Level 3: Arrival Hall
Level 2: Bus station, food garden, car parks, hotel, etc.
And the purple signage above indicates "Aras Level 2."
✅ So yes, your floor is **the 2nd floor (Level 2 / Aras 2)**.

SMEC Answer:

Based on the information in the image and a detailed analysis of the signage, your current location is indeed **3rd Floor**. Here's the complete answer and additional explanation:

1. Analysis of key information in the image:

Signages on the left:
- **5th Floor**: Balai Perlepasan (Departure Hall). This is where passengers board their planes.
- **4th Floor**: Pejabat Penyelenggara (Airline Offices). This is where airline staff work.
- **3rd Floor**: Balai Ketibaan (Arrival Hall). This is where passengers go after deplaning.
- **2nd Floor**: Hotel Sama-Sama (Hotel), Medan Kareta C & D (Parking Lot), Medan Sejera (Food Garden). These are areas offering accommodation, parking, and dining options.
- **1st Floor**: Hailing (Parking Area) and Free Shuttle to Kuala Lumpur. This is the parking and shuttle service area.

Signage on the Right:
- The sign on the right clearly states "Level 3" and has an arrow pointing in a direction. This immediately indicates that you are on the **3rd Floor**.
- The sign on the right also mentions "Stesen Bas - Terminal 2 / Destinasi Lain" (Bus Station - Terminal 2 / Other Destinations), further confirming that you are on the **3rd Floor**, as this area often houses transportation transfers.

Ground Level Signage:
- There are yellow signs on the ground that say "E-HAILING," indicating that this is a taxi or ride-hailing waiting area. These areas are usually located on lower floors (such as the 1st or 2nd floor), but the signs here are likely to help passengers quickly find the relevant services.

Pedestrians and Luggage:
- The image shows several pedestrians dragging suitcases, further confirming that you are near the **Arrival Hall**, as this is where passengers go after deplaning.

2. Expert's Additional Information:
The expert also reached the same conclusion based on the signage, stating that your current location is **3rd Floor**. The expert's analysis aligns with mine, confirming this.

3. Final Conclusion:
Based on the signage and the on-site environment, your current location is **3rd Floor**. Specifically, you are likely in or near the **Arrival Hall**, as the signage clearly states "Level 3," and this floor's function is related to the Arrival Hall.

Revised Answer:
Based on the signage in the image and the surrounding information, your current location is **3rd Floor**. The following is the specific basis for this:
1. The signage on the **left** clearly indicates the function of each floor, and **3rd Floor** is **Balai Ketibaan (Arrival Hall)**.
2. The sign on the **right** is directly labeled "Level 3" and has an arrow pointing in a certain direction, further confirming that you are on the **3rd floor**.
3. The **ground signs** and the **pedestrians carrying luggage** also match the characteristics of being near the **Arrivals Hall**.
Therefore, your current location is the **3rd floor**.

Final answer: \\boxed{3}

Q: Determine which country this restaurant is located in.

Direct Answer (from Chatgpt):

Based on the word "KELUAR" on the exit sign, this restaurant is likely located in either **Malaysia, Indonesia, Brunei, or Singapore**. "Keluar" means "Exit" in both the Malay and Indonesian languages.

SMEC Answer:

Based on clues in the image and detailed expert analysis, we can draw the following conclusions:

1. Key Clue Analysis:
- Language Sign: The sign on the door, "KELUAR," means "exit" in Malay. Malay is primarily spoken in Malaysia, Brunei, and parts of Indonesia. However, Malaysia has the most developed café culture among these countries, and the decor and ambient features in the image are highly consistent with Malaysian café culture.
- Decorative Style: The restaurant's interior features wooden tables and chairs, industrial-style pendant lights, and retro posters, creating a modern and unique yet warm design. This style is popular in Southeast Asian countries, particularly Malaysia and Singapore, particularly among independent cafés and small restaurants in cities.
- Environmental Features: The restaurant's interior features large windows, allowing for ample natural light and creating a bright and comfortable atmosphere. This design is common in Southeast Asian countries, particularly Malaysia, due to the warm and humid climates in these regions, where interior design often prioritizes natural light. Additionally, the restaurant's furniture (such as wooden tables and chairs with wrought iron frames) and air conditioning (wall-mounted) are consistent with Malaysia's climate and design practices.

2. Eliminating Other Possibilities:
Brunei: Although Brunei also speaks Malay, its smaller land area and sparse population don't allow for a more developed cafe culture than Malaysia. Therefore, Brunei is less likely. Indonesia: Indonesia's primary official language is Indonesian, not Malay. While Malay may be spoken in some areas, Indonesian cafe culture tends to be more modern and international, which doesn't align with the retro, industrial-style design in the image.

3. Additional Expert Analysis:
Based on analysis by experts on Southeast Asian cafe culture and design trends, the restaurant's wooden tables and chairs, industrial-style pendant lights, and retro posters align with cafe design trends in Malaysia and other Southeast Asian countries, particularly Malaysia's multiculturalism and its blend of modern and retro styles.
Spatial Layout and Atmosphere: The restaurant's open layout, ample use of natural light, and spacious design perfectly complement the climate and culture of Southeast Asia, particularly Malaysia.
Cultural Fusion: The restaurant's design embodies a blend of Eastern and Western elements, while respecting local culture, a typical characteristic of Malaysian cafés.
Comparison with Other Southeast Asian Countries: Singaporean cafés are generally more modern and international in design, while Thai cafés may incorporate Buddhist elements and tropical themes. This restaurant's style is more closely aligned with Malaysian characteristics.

4. Final Conclusion:
Based on the above analysis, this restaurant is most likely located in Malaysia. The Malay sign "KELUAR" is a key clue, and the restaurant's decor and surroundings are also highly consistent with Malaysian café culture..

Final answer: \\boxed{Malaysia}

Figure 11: Comparison of our method and ChatGPT on some difficult examples.

A.16 PROMPTS OF SELF-DRIVEN MULTI-EXPERT COLLABORATIVE FRAMEWORK

Meta Generation Prompt p_g

$\{question\}$
Here are your answers and those of some experts:
$\{answer\}$
Now you can create and work with multiple experts to improve your answer. So, please describe in as much detail as possible the different skills and focus you need from each expert.
We will provide each expert with the same information and queries. Each expert should have his or her own specialization covering perception, understanding and reasoning, etc., so you can assign only one subtask to each expert to ensure a more refined answer. We will relay their responses to you in turn so that you can reorganize them into better answers. Please note that descriptions should be in the second person, e.g. You are XXX.
These are the descriptions of the experts you have previously created for this task:
$\{description\}$
Therefore, do not create the same experts as above over and over again.
Now you can create a description for the new expert (please note that you can only describe one, not more than one at the same time):

Inspection Prompt p_c

$\{question\}$
We hired multiple experts to answer this question. Below is a second person description of the experts we hired: $\{existing\ description\}$
We are now hiring a new expert to help better provide the information needed for the question as well as respond to user queries. Here is a second person description of the new expert: $\{description\}$
Since there is an additional cost to hiring a new Expert, please evaluate the new Expert based on the following two criteria to decide whether or not to retain them.
1. based on the new Expert's description, determine if they can effectively assist in answering the user's question or provide the information needed for the question.
2. the new expert is not a duplicate of any existing expert.
The new expert must meet both of these criteria. If either criterion is not met, they should be discarded.
If retaining, please reply 'Retain'. If discarded, please reply: 'Discard'.

Collaboration Prompt p_c

$\{question\}$ These are you and some experts' answer: $\{answer\}$
The description of the experts you invited are: $\{description\}$
Now, you can refine your answer based on the answer and additional information they provided to better answer the question. Keep in mind that the experts' answer and additional information may not be correct, so decide carefully whether to accept his answer or stick to your original one.
Revised answer:

A.17 THE USE OF LARGE LANGUAGE MODELS (LLMS)

We have not used Large Language Models (LLMs) for our paper writing.