

MULTIVERSE: A Multi-Turn Conversation Benchmark for Evaluating Large Vision and Language Models

Young-Jun Lee [♡] Byung-Kwan Lee [♡] Jianshu Zhang [▽]
 Yechan Hwang [♡] Byungsoo Ko [♡] Han-Gyu Kim [♣] Dongyu Yao [◇] Xuankun Rong [♠]
 Eojin Joo [♡] Seung-Ho Han [♡] Bowon Ko [♡] Ho-Jin Choi [♡]

[♡] KAIST [▽] NorthwesternU [♣] NAVER [◇] CMU [♠] WHU

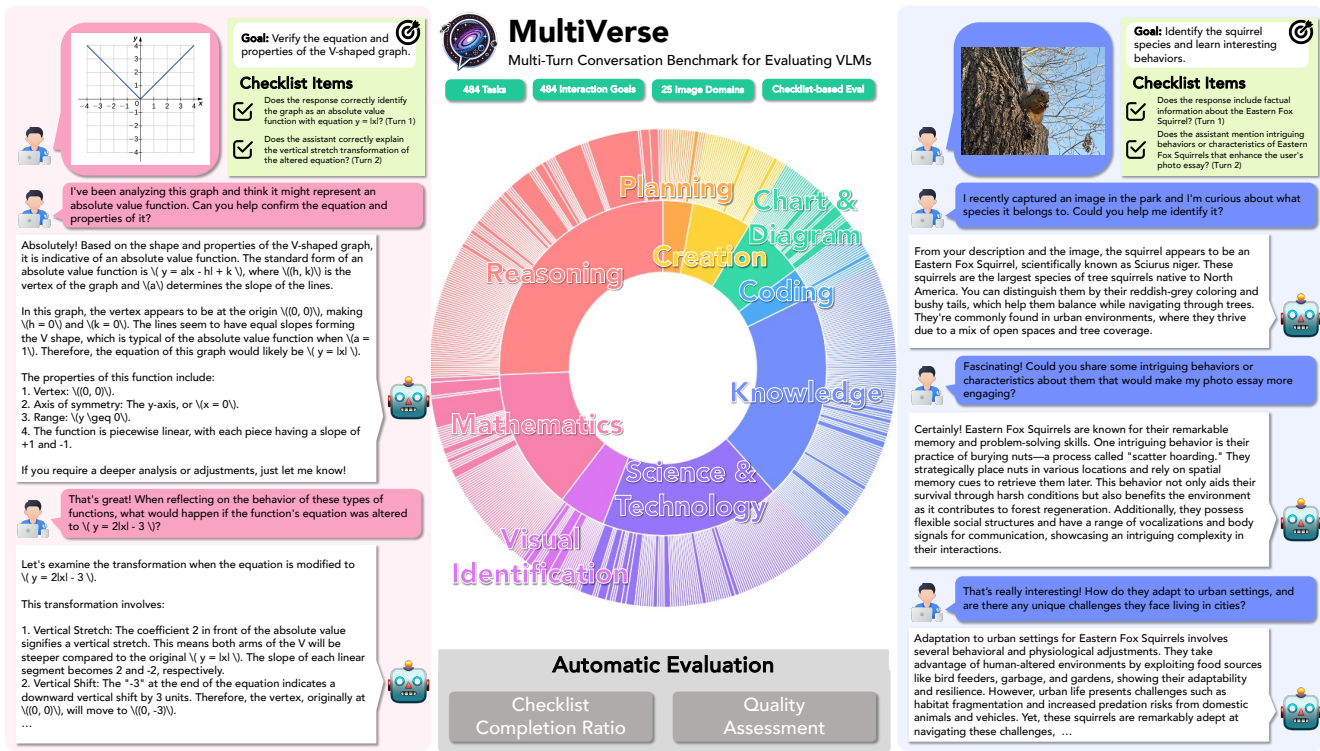


Figure 1. An overview of MULTIVERSE, a multi-turn conversation benchmark designed to evaluate VLMs across eight main tasks (e.g. Reasoning, Mathematics, Knowledge) comprising 484 tasks, nine main interaction goals (e.g. Verification, Exploration) with 484 distinct interaction goals, and 25 diverse image domains (e.g. Charts and Graphs, Nature). MULTIVERSE provides instance-specific checklist evaluation items for each turn and is the first multi-turn conversation benchmark encompassing diverse and challenging tasks.

Abstract

Vision-and-Language Models (VLMs) have shown impressive capabilities on single-turn benchmarks, yet real-world applications often demand more intricate multi-turn dialogues. Existing multi-turn datasets (e.g. MMDU, ConvBench) only partially capture the breadth and depth of conversational scenarios encountered by users. In this

work, we introduce MULTIVERSE, a novel multi-turn conversation benchmark featuring 647 dialogues—each averaging four turns—derived from a diverse set of 12 popular VLM evaluation benchmarks. With 484 tasks and 484 interaction goals, MULTIVERSE covers a wide range of topics, from factual knowledge and perception to advanced reasoning tasks such as mathematics and coding. To facilitate robust assessment, we propose a checklist-based evaluation

method that leverages GPT-4o as the automated evaluator, measuring performance across 37 key aspects, including perceptual accuracy, linguistic clarity, and factual correctness. We evaluate 18 VLMs on MULTIVERSE, revealing that even the strongest models (e.g., GPT-4o) achieve only a 50% success rate in complex multi-turn conversations, highlighting the dataset’s challenging nature. Notably, we find that providing full dialogue context significantly enhances performance for smaller or weaker models, emphasizing the importance of in-context learning. We believe MULTIVERSE is a landscape of evaluating multi-turn interaction abilities for VLMs. We make our source code and dataset publicly available.¹

1. Introduction

Vision-and-Language Models (VLMs) [3, 5, 18] have recently demonstrated remarkable performance on various evaluation benchmarks [12, 14, 37, 50, 71–74]. These benchmarks primarily assess the perception and reasoning capabilities of VLMs through simple binary, multiple-choice questions, or short free-form responses in single-turn interactions. However, real-world usage often involves users engaging in continuous, multi-turn conversations with models to solve problems iteratively. While multi-turn evaluation benchmarks have introduced for evaluating LLMs [4, 17, 26, 41, 59, 66, 78], comparatively few benchmarks (*i.e.* MMDU [47], ConvBench [45]) address VLMs’ abilities in multi-turn conversation. This gap raises the critical question: **Can VLMs that excel in single-turn benchmarks also meet user needs in more interactive, multi-turn scenarios?**

Despite recent efforts to address this issue, existing benchmarks still exhibit significant shortcomings and do not adequately cover VLMs’ capabilities in multi-turn interactions, particularly across diverse tasks (breadth) and reasoning requirements (depth). As shown in Figure 2, primarily features knowledge-oriented images (*e.g.* landscapes, animals, art) derived from WIT [60], placing substantial emphasis on knowledge-specific reasoning tasks (*e.g.* “Describe the object in the image”). Hence, MMDU offers a relatively narrow set of fewer than 15 tasks and a limited range of question styles with low lexical diversity. In contrast, ConvBench encompasses three main task categories (*i.e.* perception, reasoning, creation) and covers 219 sub-tasks derived from VisITBench [6]. This design significantly broadens its task coverage. However, it lacks more advanced tasks, such as *equation verification* in Figure 1. Furthermore, code- and mathematics-related responses remain sparse, and users’ queries are comparatively simple in their linguistic structure. These shortcomings underscore

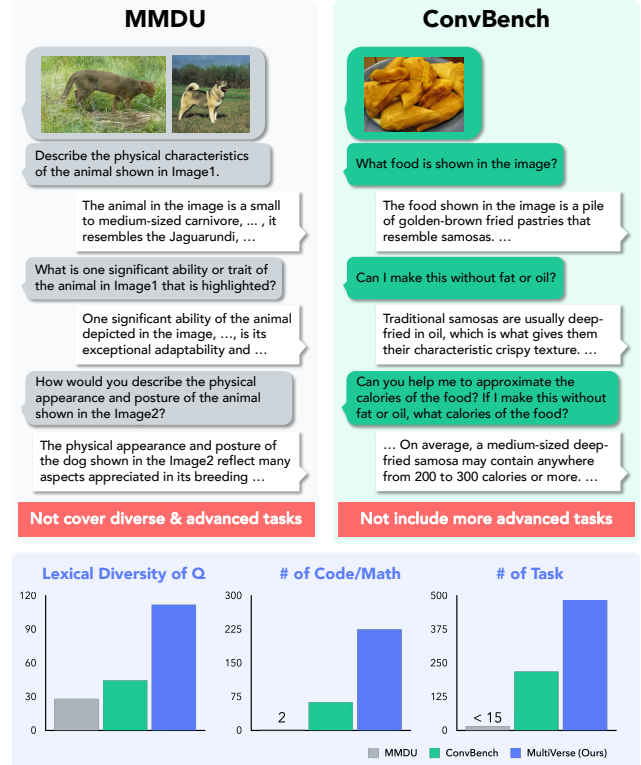


Figure 2. Examples of multi-turn conversations from MMDU [47] and ConvBench [45], along with their limitations. The graph below compares MMDU, ConvBench, and MULTIVERSE in terms of query lexical diversity, the number of code/math tasks in responses, and the total number of tasks.

the need for a more comprehensive and holistic multi-turn conversation benchmark over more diverse and challenging tasks (breadth ↑ & depth ↑).

In this paper, we introduce MULTIVERSE, a novel multi-turn conversation benchmark designed to evaluate VLMs on more diverse and advanced user tasks. MULTIVERSE comprises 647 conversations, each averaging around four turns, spanning 484 tasks (*e.g.* reasoning, mathematics) and 484 interaction goals (*e.g.* verification, analysis). We first collect source images from 12 widely used VLM evaluation benchmarks (*e.g.* MegaBench [8], CharXiv [67], MMDU [73]), covering a broad range of domains such as nature, science, and mathematics. After selecting high-quality seed images, we adopt a personal background-to-conversation approach [19, 23, 34], which ensures high lexical diversity-based multi-turn dialogues. Finally, we meticulously remove any unsuitable samples through manual review, focusing on natural conversation flow, factual correctness, and blindness criteria [15, 37]. Therefore, as shown in Figure 2, MULTIVERSE encompasses a diverse range of tasks, enabling the handling of more advanced reasoning capabilities (*e.g.* mathematics, coding) to address increasingly complex user queries with higher lexical diver-

¹<https://multiverse.github.io/>

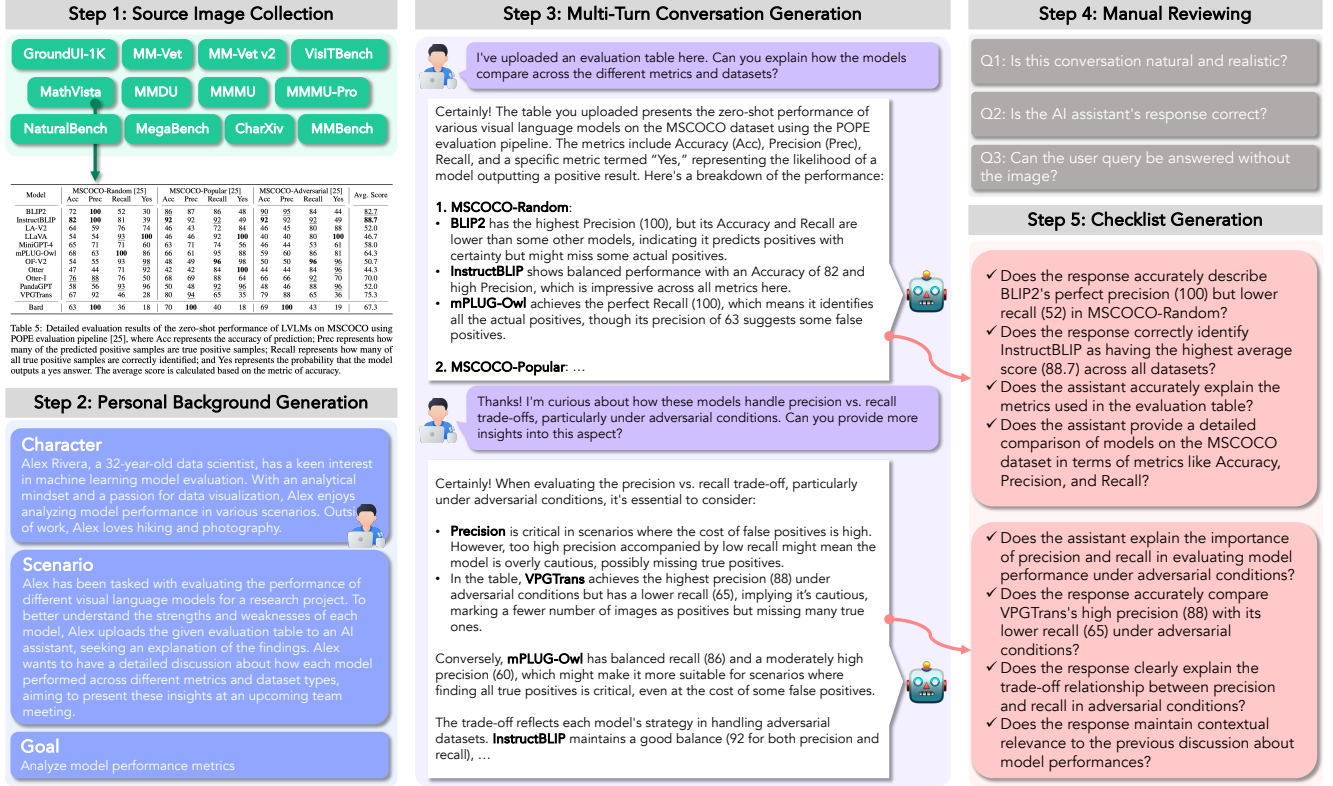


Figure 3. An overview of MULTIVERSE construction pipeline, consisting of five steps: (1) Source Image Collection, gathering images from various benchmarks; (2) Personal Background Generation, creating user personas with specific goals; (3) Multi-Turn Conversation Generation, where an AI assistant analyzes model performance based on evaluation tables; (4) Manual Reviewing, assessing response correctness and relevance; and (5) Checklist Generation, ensuring quality and coherence in AI responses.

sity. For the automatic evaluation of MULTIVERSE, we adopt a checklist-based evaluation metric inspired by prior works [33, 41], which utilize checklist items to assess open-ended generation tasks. Specifically, we employ GPT-4o as the evaluator VLM to measure response quality. Each checklist consists of multiple binary questions covering 37 key aspects, including perceptual understanding and factual correctness, as illustrated in Figure 1.

We evaluate 18 VLMs, including both open-source and proprietary models, on MULTIVERSE. Our findings reveal that most VLMs struggle with multi-turn interactions, with even the strongest model, GPT-4o, achieving less than 50% performance. This highlights a fundamental challenge in sustained dialogue reasoning. Interestingly, providing ground-truth dialogue history significantly enhances reasoning capabilities, even for smaller or weaker VLMs—demonstrating a clear in-context learning effect that enables models to adapt and refine their responses over time. Additionally, scaling up model size consistently improves multi-turn interactivity, reinforcing the importance of model capacity in complex conversational tasks. Finally, we verify that our evaluation metrics remain robust against verbosity bias, ensuring reliable and fair assessment across

models. These insights shed light on the limitations of current VLMs and underscore the necessity of improving contextual understanding in long-form interactions.

2. MULTIVERSE

In this section, we will describe how we construct MULTIVERSE, with an overview shown in Figure 3.

2.1. Dataset Construction

Step 1: Source Image Collection. We began by collecting source images spanning diverse topics (e.g. science, mathematics, and natural scenes) from 12 existing evaluation benchmarks: MMDU [47], GroundUI-1K [79], MMMU [73], MMMU-Pro [74], NaturalBench [37], VisIT-Bench [6], MathVista [50], MM-Vet [71], MM-Vet v2 [72], CharXiv [67], MMBench (Eng) [46], and MegaBench [8]. In total, 49,700 images were obtained from the test sets of these benchmarks. We then applied a rigorous three-step data refinement process to ensure high-quality selections: (1) **Deduplication:** We used pHash to detect and remove duplicate images, resulting in 13,998 images (29.1%) being discarded; (2) **Image Quality Scoring:** We employed GPT-4o-2024-11-20 to rate each image on a 1–5 scale in terms of

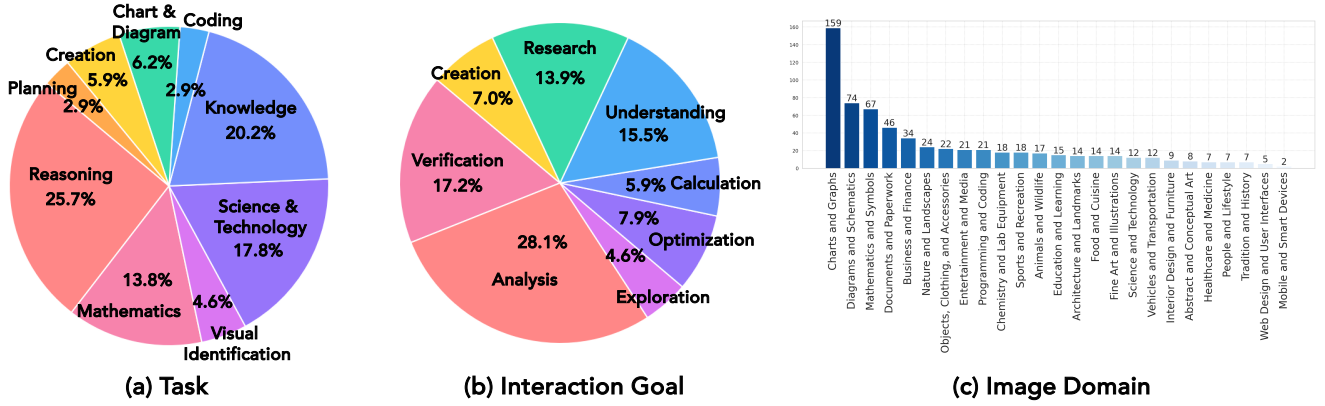


Figure 4. Detailed distribution of (a) Task, (b) Interaction Goal, and (c) Image Domain in MULTIVERSE.

clarity, resolution, and real-world plausibility. Images scoring below 5 in any category were removed (21,808 images, 64.77%). See the Appendix for examples of excluded low-quality images; (3) **Image Category Classification**: We next asked GPT-4o to classify the category of each image, resulting in 57 categories obtained. The most common categories were *Charts and Graphs* (22.92%) and *Diagrams and Schematics* (11.17%). Categories with fewer than 50 images were excluded (213 images, 1.80%); (4) **Weighted Image Category Sampling**: To ensure our benchmark remains compact and accessible for other researchers to evaluate without excessive computational burden, we maintained a relatively small scale. Accordingly, we performed a weighted random sampling of images from each category, proportional to the overall category distribution, while capping the total number of selected images at 1K.

Step 2: Personal Background Generation. To construct more plausible and realistic multi-turn conversations, we draw inspiration from real-world scenarios. Most people typically engage in dialogue with a specific goal in mind—such as resolving a problem by providing an image alongside a relevant query—and these problems often relate to the user’s situational context and personal background (e.g. age, profession, hobbies). Hence, before generating the multi-turn conversation based solely on the image, we first create a *fictional character* and *scenario context* to represent a multi-turn exchange between this character and an AI assistant, along with a *goal* tied to the scenario, using GPT-4o.² Leveraging the *personal background-to-conversation* approach [19, 23, 34] enables the construction of multi-turn conversation datasets with LLMs, leading to richer, more diverse, and higher-quality dialogues.

²In pilot experiments, we observed that generating multi-turn conversations solely from an image led to less diverse and lower-quality outcomes.

Step 3: Multi-turn Conversations Generation. Building on the fictional character, scenario context, and defined goal, we use GPT-4o to generate multi-turn conversations between the character and the AI assistant. These conversations follow four key principles:

- **Detailed & Informative Responses:** The assistant’s responses must be detailed, specific, expert-level, and highly informative.
- **Increasing Complexity:** Each subsequent user query becomes progressively more challenging, creative, and complex, as motivated by prior work [45].
- **Diverse Linguistic Styles:** To emulate realistic human interactions, user utterances vary in linguistic style instead of relying on task-specific prompts commonly seen in existing benchmarks [47] (e.g. “describe,” “what is,” “where is”).
- **Conversation Length:** Each conversation consists of four turns, yielding a total of eight utterances (four from the user and four from the assistant).

In total, we obtained 839 multi-turn conversations after removing 161 samples due to degeneracy issues with GPT-4o, such as parsing failures or insufficient turn length (<6).

Step 4: Manual Reviewing. To build a realistic and rigorous multi-turn conversation benchmark, we manually review the generated conversations according to three primary criteria:

- **Naturalness & Realism:** We remove any conversation that does not flow realistically or contains user queries that are implausible given the image. For example, if the image is a complex scatter plot with vibrant color dots, yet the user’s query solely focuses on the color rather than the actual content of the plot, we deem it unrealistic and remove the sample.
- **Correctness:** We conservatively remove conversations if the AI assistant’s response is clearly incorrect or provides misleading information.

Statistic	Number
Total Conversations	647
Unique number of images	647
Avg./Max. number of turns	3.91/4
Interaction goal/sub-goal classes	8/484
Image main/sub classes	25/384
Task main/sub classes	9/484
Avg./Max. question length	30.53/70
Avg./Max. answer length	221.51/742
Avg./Max. question diversity	111.97/235.48
Avg./Max. answer diversity	117.96/230.48
Unique number of checklist items	21995
Unique number of main/sub key aspect	37/10115

Table 1. Statistics of MULTIVERSE. Lexical diversity of questions and answers is measured using MTLT [52], while question and answer length is computed with the LLaMA-3.1-8B tokenizer.

- **Blindness:** Inspired by prior works [12, 37] addressing the “blind” issue, we remove conversations in which the AI assistant can sufficiently solve the user’s query without referencing the provided image.

Following this process, we retained 647 conversations and removed 192. Specifically, 48 were removed for unnatural flow, 65 for incorrect responses, and 104 for blindness-related issues, with some overlap across these criteria.

2.2. Analysis of MULTIVERSE

Basic Statistics. As shown in Table 1, MULTIVERSE consists of 647 conversations, averaging 3.91 turns per conversation. It spans 8 primary interaction goals and 484 sub-interaction goals, allowing for a comprehensive evaluation of VLMs’ ability to handle diverse and realistic user interactions. Additionally, MULTIVERSE covers nine main tasks with 484 corresponding sub-tasks. MULTIVERSE exhibits high lexical diversity in both queries and responses, suggesting that it necessitates advanced reasoning to effectively address complex user queries.

Detailed Distributions. Figure 4 presents the detailed distribution of (a) tasks, (b) interaction goals, and (c) image domains in MULTIVERSE. (a) MULTIVERSE exhibits a high proportion of reasoning tasks, followed by knowledge-related tasks. Compared to MMDU and ConvBench, MULTIVERSE includes more advanced reasoning tasks (e.g., mathematics, charts and diagrams, coding). (b) MULTIVERSE highlights the diversity of interaction goals in MULTIVERSE, where “Analysis” is the primary motivation for multi-turn interactions in MULTIVERSE, followed by “Verification”, both of which reflect realistic and practical interaction scenarios. (c) MULTIVERSE contains a substantial number of image domains, with “Charts and Graphs”

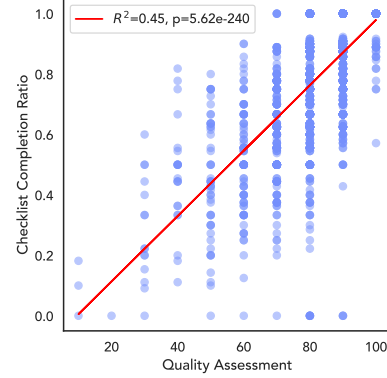


Figure 5. Correlation between the checklist completion ratio and quality assessment. The red line represents the best-fit linear regression, indicating a strong positive correlation ($R^2 = 0.44$) between the two metrics.

being the most predominant, followed by “Diagrams and Schematics”. These distributions demonstrate MULTIVERSE’s capacity to comprehensively evaluate VLM capabilities across diverse and challenging tasks, realistic interaction goals, and a broad spectrum of image types.

2.3. Automatic Evaluation Process

Recently, the use of powerful LLMs/VLMs (e.g., GPT-4o) as evaluators has become a standard approach for assessing response quality in open-ended tasks [24, 25, 32, 78]. This is typically done by assigning an integer score between 1 and 10 to measure response quality. While this method is both straightforward and effective, recent studies [33, 41] have raised concerns about its robustness. Therefore, checklist-based evaluation methods have been introduced to improve reliability and interpretability. In this work, we adopt an instance-specific, checklist-based evaluation approach to provide more interpretable and robust evaluation results. Specifically, we generate a unique checklist for each user query at every turn in a multi-turn conversation. Each checklist consists of multiple binary questions, with each question focusing on a specific key aspect (e.g. perception, factual correctness). To ensure quality, we generate these checklists using GPT-4o and Claude-3.5-Sonnet, followed by a manual review conducted by the authors to validate their accuracy and reliability.

Evaluation Metrics. We employ GPT-4o-2024-11-20 as the evaluator VLM, prompting it to assess the quality of generated responses at every turn in a multi-turn conversation based on a checklist and a reference answer³. The

³Existing studies have demonstrated that providing a reference answer improves the correlation between human and model judgments. Therefore, we incorporate a reference answer in our evaluation.

Models	Turn 1	Turn 2	Turn 3	Turn 4	Avg.	r
InternVL2.5-1B	13.93	21.36	23.51	26.08	21.22	3.86
InternVL2.5-2B	17.79	28.96	30.80	34.32	27.97	5.14
InternVL2.5-4B	27.14	37.24	38.79	39.46	35.66	3.85
Qwen2.5-VL-3B	32.80	37.41	38.55	44.05	38.20	3.49
LLaVA-1.5-7B	9.10	26.43	29.14	31.81	24.12	7.08
LLaVA-NeXT-7B	13.59	28.25	29.94	33.07	26.21	6.01
LLaVA-OneVision-7B	24.83	34.47	37.72	37.81	33.71	4.22
Qwen2-VL-7B	35.93	37.99	39.35	43.10	39.09	2.29
Qwen2.5-VL-7B	45.13	47.60	49.31	50.58	48.15	1.81
InternVL2-8B	32.09	39.37	38.98	38.88	37.33	2.00
InternVL2.5-8B	22.90	37.85	38.89	42.15	35.45	5.88
LLaMA-3.2-11B-Vision	14.38	20.02	25.41	26.63	21.61	4.21
Qwen2.5-VL-72B	52.05	47.72	45.82	46.19	47.95	-1.95
LLaMA-3.2-90B-Vision	24.92	37.63	38.99	41.58	35.78	5.13
Gemini-2.0-Flash	42.03	49.37	51.23	48.41	47.76	2.10
Claude-3.5-Sonnet	46.60	47.16	48.30	45.00	46.76	-0.37
Claude-3.7-Sonnet	43.38	44.39	44.77	43.70	44.06	0.13
GPT-4o	48.56	50.28	50.54	49.12	49.63	0.19

Table 2. Average performance of 18 VLMs across multiple turns in MULTIVERSE under the *Oracle* setting. r represents the slope of the performance increase as interactions progress.

prompt template used for evaluation is provided in the Appendix. Intuitively, we posit that a high-quality response should successfully satisfy the checklist items while achieving a higher integer score on a 1 to 10 scale. Based on this premise, our metric consists of two sub-metrics:

- **Checklist Completion Ratio:** This measures how well the generated response addresses the given checklist items by calculating the ratio of “Yes”.
- **Quality Assessment:** This assigns an integer score between 1 and 10 to evaluate the overall quality of the response. The score is then scaled up from 1–10 to 10–100 for better interpretability.

As shown in Figure 5, Checklist Completion Ratio and Quality Assessment exhibit a strong positive correlation, indicating that responses meeting more checklist items tend to receive higher quality scores. Based on this, our final score is computed as their product.

3. Experiments

3.1. Experimental Setup

Evaluated Models. We evaluate in total 18 VLMs, categorized into: (i) Proprietary models: GPT-4o-2024-11-20 [18], Claude-3.5-Sonnet-2024-10-22 [3], Claude-3.7-Sonnet-2025-02-19 [2], Gemini-2.0-Flash-001 [16]; (ii) Open-source models: LLaVA-1.5-7B [43], LLaVA-NeXT-7B [42], LLaVA-OneVision-7B [38], Qwen2-VL-7B [64], Qwen2.5-VL-{3, 7, 72}B-Instruct [5], LLaMA-3.2-{11, 90}B-Vision-Instruct [53], InternVL2-8B [10],

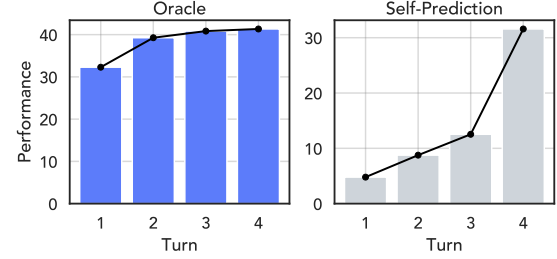


Figure 6. Average performance of 18 VLMs under two different evaluation settings: *Oracle* and *Self-Prediction*.

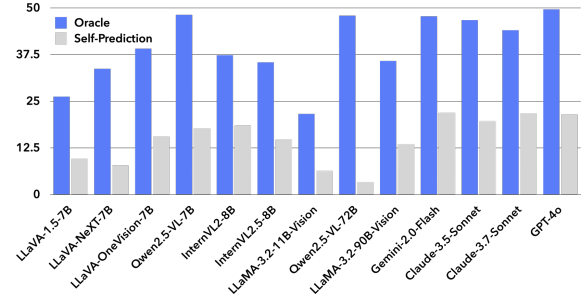


Figure 7. Comparison performance across 19 different VLMs between *Oracle* and *Self-Prediction* settings.

InternVL2.5-{1,2,4,8}B [11].

Evaluation Settings. Since MULTIVERSE is a multi-turn conversational benchmark, the dialogue history significantly impacts response quality. To ensure a fair comparison of different VLMs, we adopt the ground-truth dialogue history provided by MULTIVERSE as the default evaluation setting, **Oracle**. This approach is widely used in existing multi-turn conversation benchmarks [4, 45]. However, relying solely on ground-truth dialogue history may introduce biases, as variations in linguistic style and model capacity can disproportionately affect performance across different VLMs. Therefore, we additionally assess VLMs under the **Self-Prediction** setting, in which the model generates its own dialogue history.

3.2. Experimental Results and Analysis

Powerful VLMs still struggle with multi-turn interactions. As shown in Table 2, all VLMs exhibit relatively low performance ($< 50\%$), indicating that multi-turn interactions in MULTIVERSE remain challenging, even for high-performing VLMs in the *Oracle* setting. GPT-4o achieves the highest average performance. Overall, proprietary VLMs generally outperform open-source VLMs, with the exception of the Qwen2.5-VL series, which notably surpasses Claude-3.5-Sonnet, Claude-3.7-Sonnet, and Gemini-2.0-Flash. Additionally, Claude-3.7-Sonnet scores 2.7% lower than Claude-3.5-Sonnet. Among open-source

models, LLaMA-3.2-11B-Vision achieves the lowest performance, performing even worse than LLaVA-1.5-7B.

As interaction progresses, providing ground-truth dialogue context unlocks the reasoning capabilities. As shown in Figure 6, overall performance steadily improves across multiple turns in the Oracle setting. This result suggests that incorporating the golden dialogue history helps VLMs better understand contextual nuances, serving as a guiding mechanism for resolving user queries in long-term conversations. By leveraging the golden dialogue, even smaller or weaker VLMs can adapt to similar linguistic styles and phrases from context, leading to improved performance. Further analysis in Table 2 presents the performance improvement of each model across multi-turn interactions by measuring the slope (r). Overall, most VLMs benefit from extended interactions, demonstrating the effects of in-context learning. Interestingly, smaller or weaker VLMs experience the most significant gains, indicating that the ground-truth dialogue history in MULTIVERSE helps unlock their reasoning capabilities by providing strong contextual guidance. However, Qwen2.5-VL-72B and Claude-3.5-Sonnet do not show improvement when using the ground-truth dialogue history. This negative slope may be attributed to linguistic differences between these models and GPT-4o, which serves as the reference dialogue in MULTIVERSE.

Oracle vs. Self-Prediction. Figure 7 compares the performance of VLMs under Oracle and Self-Prediction. Overall, models exhibit significantly improved responses in the Oracle setting compared to Self-Prediction that relies on self-generated dialogue history, leading to a substantial performance gap (with a maximum drop of -44.64% observed in Qwen2.5-VL-72B). These findings indicate that providing ground-truth dialogue history serves as a form of in-context learning, effectively guiding models to generate responses that are more closely aligned with reference answers. Interestingly, Qwen2.5-VL-Series models (7B and 72B) demonstrate a notably larger performance gap compared to other VLMs (-30.44 for 7B, -44.64 for 72B). This suggests that the backbone language models of Qwen2.5-VL and GPT-4o may exhibit distinct linguistic styles or response patterns. These results underscore the critical role of supplying precise and coherent dialogue context to enhance overall model performance.

VLMs demonstrate selective interactivity capabilities. As shown in Table 3, most VLMs exhibit strong performance in analysis (🔍) and understanding (💡). However, their performance in optimization (⚙️) and research (🔬) tasks—typically demanding innovative thinking—remains

Models	📊	🔍	🔬	⚙️	💡	🔬	🔭
InternVL2.5-1B	20.37	24.55	23.98	15.06	16.95	22.03	18.36
InternVL2.5-2B	26.67	31.48	24.82	22.74	26.23	28.70	26.51
InternVL2.5-4B	34.43	38.56	32.14	31.56	34.46	40.90	29.65
Qwen2.5-VL-3B	36.08	43.19	34.56	29.43	37.61	43.46	35.91
LLaVA-1.5-7B	20.34	26.58	26.27	19.57	14.11	23.32	28.70
LLaVA-NeXT-7B	21.78	30.36	28.62	21.75	14.26	25.41	29.06
LLaVA-OneVision-7B	30.55	36.29	31.58	27.97	31.84	37.27	31.73
Qwen2-VL-7B	36.28	43.66	35.95	31.62	35.77	43.54	35.79
Qwen2.5-VL-7B	46.14	54.23	44.23	40.64	44.86	54.71	39.73
InternVL2-8B	34.85	42.06	33.00	28.76	33.74	41.76	33.19
InternVL2.5-8B	33.55	38.39	32.95	29.92	34.45	39.58	30.64
LLaMA-3.2-11B-Vision	17.94	23.30	20.43	18.84	16.51	22.48	23.41
Qwen2.5-VL-72B	45.58	52.10	42.86	42.59	47.15	55.40	43.32
LLaMA-3.2-90B-Vision	33.96	38.19	33.41	32.17	34.68	39.03	32.57
Gemini-2.0-Flash	43.68	52.25	40.73	44.46	48.92	54.48	42.58
Claude-3.5-Sonnet	46.15	50.94	43.80	40.43	47.72	53.95	37.99
Claude-3.7-Sonnet	42.58	47.43	43.18	38.83	45.33	51.22	38.56
GPT-4o	46.80	53.67	42.70	46.36	50.54	56.41	43.50

Table 3. Performance of 18 VLMs in MULTIVERSE for 8 different interaction goals—verification (📊), analysis (🔍), exploration (🔬), optimization (⚙️), calculation (💡), understanding (💡), research (🔬), and creation (🔭)—under Oracle setting.

comparatively weak. Among these, the Qwen2.5-VL series demonstrates relatively stronger interactivity across all interaction goals. On the other hand, weaker or smaller VLMs with lower performance, such as LLaVA-1.5-7B and LLaVA-NeXT-7B, struggle particularly with verifiable tasks, including optimization (⚙️) and calculation (💡).

Effect of Model Scaling. Figure 8 presents performance comparisons across varying model sizes in three different VLM groups: InternVL2.5, LLaMA-3.2-Vision, and Qwen2.5-VL. In general, larger models exhibit higher overall performance on most tasks. However, within the Qwen2.5-VL series, increasing the model size does not always guarantee improvements, and its impact appears task-dependent. For instance, Qwen2.5-VL-72B achieves stronger reasoning on verifiable tasks (e.g. mathematics, coding, and interpreting charts & diagrams), while Qwen2.5-VL-7B shows advantages in creative thinking. These findings suggest that larger models generally offer enhanced interactivity capabilities, although the benefit may vary based on the specific task.

Does automatic evaluation metric possess verbosity bias? We explore whether our evaluation metric favors longer answers across multi-turns. As shown in Figure 9, the linear correlation (R^2) between length of response and performance from GPT-4o is weakened as interaction increases (until turn 3). As interaction progress until turn 3, the p -value increases and R -value decreases, which suggests that our evaluation metric could reduce verbosity bias and length dependency.

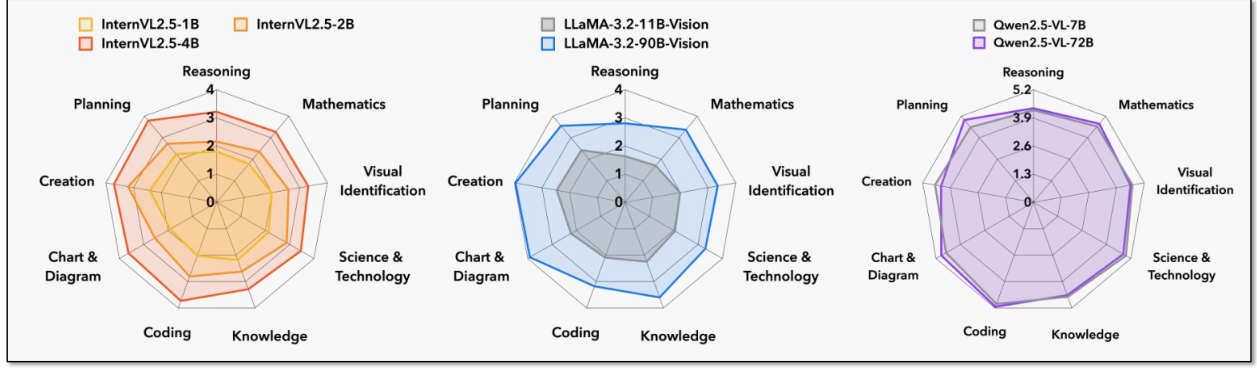


Figure 8. Performance comparison across different model sizes in three model families: InternVL2.5, LLaMA-3.2-Vision, and Qwen2.5-VL. The radar charts show that larger models generally perform better across tasks, but scaling effects vary. Notably, Qwen2.5-VL-72B excels in structured reasoning tasks (e.g., mathematics, coding), while Qwen2.5-VL-7B shows stronger creative abilities, highlighting task-dependent scaling impacts.

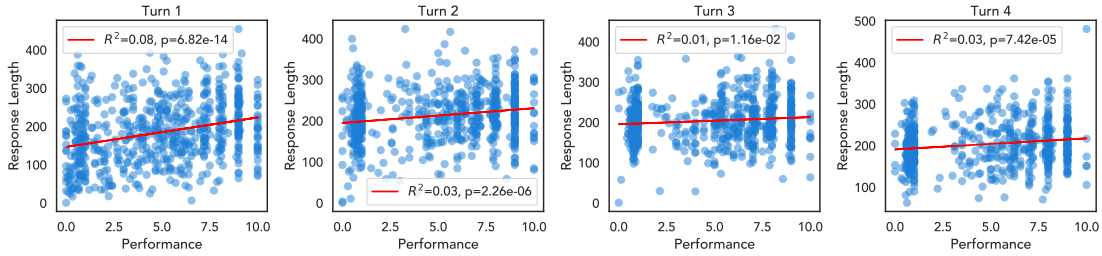


Figure 9. Linear correlation (R^2) between response length and performance from GPT-4o.

4. Related Work

Single-Turn Evaluation Benchmarks. To comprehensively evaluate multiple vision-language capabilities of recent powerful VLMs, many benchmarks have been introduced. These benchmarks assess integrated capabilities (*i.e.* perception and reasoning) [9, 14, 35, 36, 46, 56, 68, 71–74, 77], text-richness (*e.g.* charts, diagrams) [22, 51], mathematical reasoning [21, 50, 76], scientific understanding [49], perception-specialized tasks that are easy for humans to perceive [15, 37, 55, 57, 62, 63, 75], and web & mobile applications [44]. Despite their breadth, most existing benchmarks rely on single-turn evaluations—typically in the form of binary, multiple-choice, or free-form responses—thus failing to capture the effectiveness of VLMs in multi-turn interaction. This gap highlights the need for a new benchmark that better reflects practical, real-world conversational scenarios.

Multi-Turn Evaluation Benchmarks. In the NLP domain, numerous multi-turn conversation benchmarks have recently been introduced to assess the interactive capabilities of LLMs [4, 17, 26, 41, 59, 66, 78]. By contrast, only a few benchmarks exist for evaluating the multi-turn interaction capabilities of VLMs, such as MMDU [47] and Con-

vBench [45]. However, these existing benchmarks still fall short of assessing VLM performance in real-world human-AI interactions. In particular, they do not address more advanced and complex reasoning tasks (*e.g.* verifying equations shown in the image), and most queries can be resolved through straightforward statements rather than more elaborate formats, such as code or mathematical expressions. These limitations hinder a comprehensive evaluation of VLMs under a wide range of scenarios, underscoring the need for a richer, more rigorous multi-turn conversation benchmark.

5. Conclusion

In this work, we introduce MULTIVERSE, a new multi-turn conversation benchmark designed to evaluate VLMs. Our dataset encompasses a diverse range of interaction goals (*e.g.* verification, analysis) and tasks (*e.g.* reasoning, mathematics), providing a realistic and challenging testbed for assessing the multi-turn interactive capabilities of VLMs. Moreover, we propose an instance-specific, checklist-based evaluation framework comprising two key sub-metrics: checklist completion ratio and quality assessment. We reveal that even recent state-of-the-art VLMs struggle with handling multi-turn interactions effectively.

Acknowledgement

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-II220641, XVoice: Multi-Modal Voice Meta Learning)

References

- [1] Anthropic. The claude 3 model family: Opus, sonnet, haiku. <https://www.anthropic.com>, 2024. 14
- [2] Anthropic. Claude 3.7 sonnet system card. 2025. 6
- [3] AI Anthropic. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3(6), 2024. 2, 6
- [4] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*, 2024. 2, 6, 8
- [5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6
- [6] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 2, 3, 14
- [7] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827, 2024. 14
- [8] Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyang Jiang, Bohan Lyu, et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*, 2024. 2, 3, 14
- [9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 8
- [10] Zhe Chen, Jiannan Wu, Wenhao Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 6
- [11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6
- [12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2, 5
- [13] et al. Erfei Cui. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2024. <https://sharegpt4o.github.io/>. 14
- [14] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2, 8
- [15] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 2, 8
- [16] Gemini. Gemini 2.0 flash. 2024. 6
- [17] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024. 2, 8
- [18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 6, 14
- [19] Jihyoung Jang, Minseong Boo, and Hyoungun Kim. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. *arXiv preprint arXiv:2310.13420*, 2023. 2, 4
- [20] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. Enhancing multimodal large language models with vision detection models: An empirical study. *arXiv preprint arXiv:2401.17981*, 2024. 14
- [21] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s up” with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 8
- [22] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 8
- [23] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*, 2022. 2, 4
- [24] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*, 2023. 5

- [25] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models, 2024. 5
- [26] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*, 2024. 2, 8
- [27] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 14
- [28] Byung-Kwan Lee, Sangyun Chung, Chae Won Kim, Beomchan Park, and Yong Man Ro. Trol: Traversal of layers for large language and vision models. *arXiv preprint arXiv:2406.12246*, 2024. 14
- [29] Byung-Kwan Lee, Chae Won Kim, Beomchan Park, and Yong Man Ro. Meteor: Mamba-based traversal of rationale for large language and vision models. *arXiv preprint arXiv:2405.15574*, 2024. 14
- [30] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Collavo: Crayon large language and vision model. *arXiv preprint arXiv:2402.11248*, 2024. 14
- [31] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Moai: Mixture of all intelligence for large language and vision models. *arXiv preprint arXiv:2403.07508*, 2024. 14
- [32] Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation, 2024. 5
- [33] Yookyung Lee, Joonghoon Kim, Jaehye Kim, Hyowon Cho, and Pilsung Kang. Checkeval: Robust evaluation framework using large language model via checklist. *arXiv preprint arXiv:2403.18771*, 2024. 3, 5
- [34] Young-Jun Lee, Dokyong Lee, Junyoung Youn, Kyeongjin Oh, Byungsoo Ko, Jonghwan Hyeon, and Ho-Jin Choi. Stark: Social long-term multi-modal conversation with persona commonsense knowledge. *arXiv preprint arXiv:2407.03958*, 2024. 2, 4
- [35] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 8
- [36] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 8
- [37] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024. 2, 3, 5, 8, 14
- [38] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 14
- [39] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 14
- [40] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 14
- [41] Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024. 2, 3, 5, 8
- [42] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6, 14
- [43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 6
- [44] Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*, 2024. 8
- [45] Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194*, 2024. 2, 4, 6, 8
- [46] 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? *arXiv preprint arXiv:2307.06281*, 2023. 3, 8, 14
- [47] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024. 2, 3, 4, 8, 14
- [48] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 14
- [49] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 8
- [50] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathemat-

- ical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2, 3, 8, 14
- [51] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 8
- [52] Philip M McCarthy and Scott Jarvis. Mtl-d, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010. 5
- [53] Meta. Llama 3.2 vision. 2024. 6
- [54] OpenAI. Gpt-4v(ision) technical work and authors, 2023. <https://openai.com/contributions/gpt-4v>, Last accessed on 2024-02-13. 14
- [55] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized visual instruction tuning. *arXiv preprint arXiv:2410.07113*, 2024. 8
- [56] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions. *arXiv preprint arXiv:2406.07502*, 2024. 8
- [57] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024. 8
- [58] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 14
- [59] Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritiz, Willow Primack, Summer Yue, and Chen Xing. Multi-challenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *arXiv preprint arXiv:2501.17399*, 2025. 2, 8
- [60] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021. 2
- [61] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 14
- [62] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 8
- [63] 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*, pages 9568–9578, 2024. 8, 14
- [64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6
- [65] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 14
- [66] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*, 2023. 2, 8
- [67] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024. 2, 3, 14
- [68] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 8
- [69] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 14
- [70] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 14
- [71] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2, 3, 8, 14
- [72] Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024. 3, 14
- [73] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 2, 3, 14
- [74] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 2, 3, 8, 14

- [75] Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. *arXiv preprint arXiv:2502.12084*, 2025. [8](#)
- [76] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. [8](#)
- [77] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qing-song Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. [8](#)
- [78] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. [2](#), [5](#), [8](#)
- [79] Longtao Zheng, Zhiyuan Huang, Zhenghai Xue, Xinrun Wang, Bo An, and Shuicheng Yan. Agentstudio: A toolkit for building general virtual agents. *arXiv preprint arXiv:2403.17918*, 2024. [3](#), [14](#)