Seeing is believing: Comprehensive Self-Reflective Evaluation System for Large Multi-modal Models

Anonymous ACL submission

Abstract

The rapid advancement of large multi-modal models has generated an immediate demand for comprehensive evaluation methodologies. In this paper, we introduce a novel and systematic *Self-Reflective Evaluation System* (SRES) framework for comprehensive multi-modal model evaluation. Unlike traditional frameworks, our SRES uniquely integrates three core dimensions (Visual, Linguistic, and Robustness) to comprehensively cover evaluation tasks while enabling synchronized multi-dimensional assessment for holistic multi-modal analysis. Importantly, we establish the first standardized dynamic assessment mechanism by incorporating a novel self-reflective module, which autonomously assesses performance and conducts process optimization without human intervention. Additionally, we construct a comprehensive benchmark dataset comprising 352 subtasks to systematically evaluate 15 leading large multi-modal models. Through rigorous multi-dimensional comparative analysis, we assess their performance metrics and robustness characteristics. The framework implementation and benchmark data are publicly available at: https://anonymous.4open.science/r/SRES-B2B

1 Introduction

Large Multi-modal Models (LMMs) have achieved significant advancements in recent years, with numerous architectures demonstrating effectiveness across diverse domains (Dai et al., 2023; Zhu et al., 2024; Li et al., 2023). Nevertheless, the community lacks a standardized benchmarking framework to systematically evaluate their holistic capabilities (Liu et al., 2024b; Yu et al., 2024; Liu et al., 2024c; Schwenk et al., 2022).

As shown in Figure 1, the current LMM evaluation system (Singh et al., 2019; Guetta et al., 2023; Du et al., 2024; Li et al., 2024; Liu et al., 2024a) narrowly focus on singular evaluation metrics that fail to capture the multi-dimensional complexity required for real-world applications. This can be further elaborated in the following three fundamental aspects:

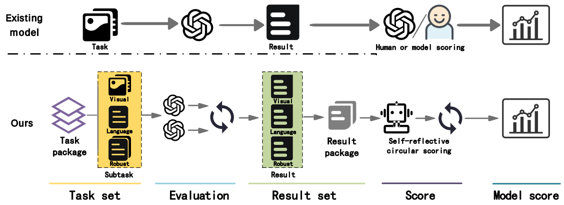


Figure. 1. Comparison of our proposed SRES with the existing mainstream evaluation system. Task refers to a single question in the input model.

1) Modular evaluation frameworks fail to address inherent model architectural limitations. Mainstream LMMs typically integrate a visual translator alongside the core Large Language Model (LLM) through partitioned designs (Verma et al., 2024), yet this structural segregation introduces critical evaluation blind spots. By restricting visual processing to the translator's domain-specific capacities (*e.g.*, image-to-text conversion) while confining linguistic reasoning within LLMs' pre-trained syntactic boundaries (Goyal et al., 2017), existing frameworks systematically handle visual and linguistic components as discrete modules while neglecting intrinsic multi-modal integration (Guetta et al., 2023; Saikh et al., 2022; Nemani et al., 2023; Xu et al., 2023).

2) Robustness evaluation inadvertently obscures assessment results. Existing evaluation frameworks typically rely on repeated experimental trials to calculate performance averages. While this approach is widely adopted, it often incurs substantial computational overhead and fails to quantify system robustness adequately. The stability of a model across multiple runs should be a core evaluation criterion for accurately assessing its true

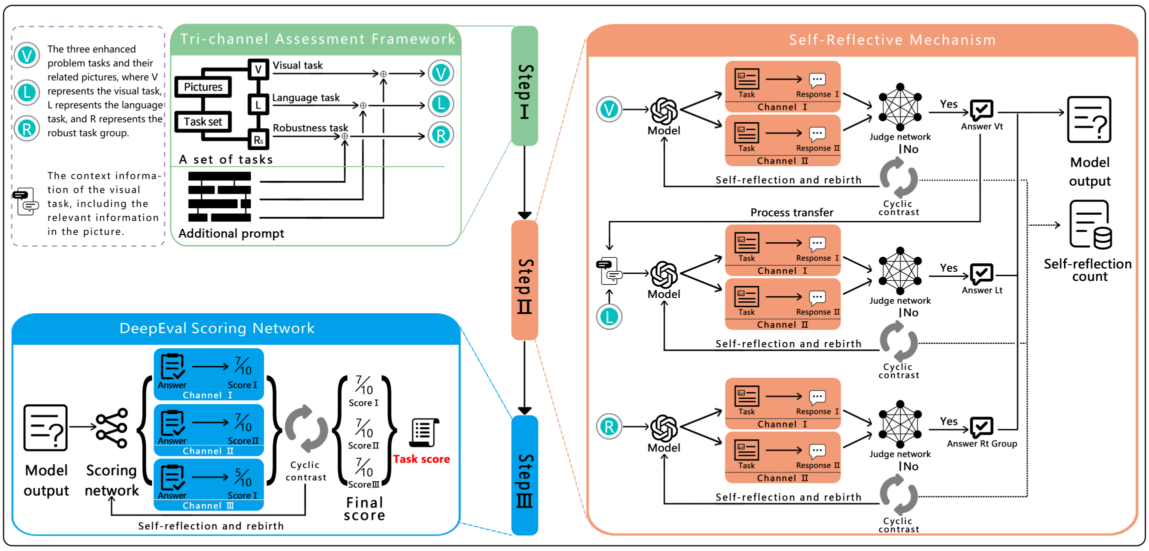


Figure. 2. Our framework implements a three-stage evaluation pipeline: (I) Tri-channel Assessment Framework, (II) Self-Reflective Mechanism, and (III) DeepEval Scoring Network, assessing visual, linguistic, and robustness capabilities.

capabilities. However, most evaluation systems neglect this dimension, treating robustness as an afterthought rather than a foundational requirement. This oversight can lead to misleading evaluations by conflating model instability with inherent capability limitations (Sun et al., 2024; Lou et al., 2023; Jing et al., 2024).

3) Task monotony disconnected from real-world complexity. The evaluation and task design of models face significant challenges due to the diverse array of scenarios, which encompass both fixed-format responses and open-ended inquiries. Current systems exhibit critical limitations in addressing real-world complexity, primarily because of their siloed architectural designs and fragmented assessment methodologies. These systems rely on a task-isolated paradigm, necessitating separate benchmark executions for each scenario, which is fundamentally unscalable as the number of scenarios grows. Moreover, the lack of task diversity in assessment design undermines cross-domain adaptability, undermining the reliability of evaluations in real-world applications.

Taking into account these limitations, we propose the *Self-Reflective Evaluation System* (SRES) to allow a comprehensive evaluation of the capabilities of LMMs. The main contributions of this paper are:

- We establish a comprehensive evaluation system that seamlessly integrates visual comprehension, linguistic understanding, and robustness testing within a unified architecture. The framework captures intrinsic multi-modal interactions through carefully constructed task inter-dependencies, enabling comprehensive capability evaluation while eliminating human bias and ensuring reproducibility.

- To address the critical challenge of output variability in model assessments, we introduce a novel self-reflective module designed to dynamically adjust and mitigate output fluctuations, thereby substantially enhancing the reliability of assessments.

- We develop a meticulously designed evaluation dataset and conduct a comprehensive comparison of 15 state-of-the-art mainstream LLMs, offering the most comprehensive performance analysis.

2 SRES

As illustrated in Figure 2, our SRES framework implements a three-stage evaluation pipeline:

Step I: We first design a tri-channel assessment framework to simultaneously evaluate three core dimensions: Visual (V) comprehension, Linguistic (L) understanding, and Robustness (R) testing.

Step II: The self-reflective module employs iterative introspective reasoning, thereby enabling accurate performance evaluation.

Step III: Finally, we develop a comprehensive DeepEval scoring network powered by advanced DeepSeek-R1 models to quantitatively obtain accurate scores for each LMM.

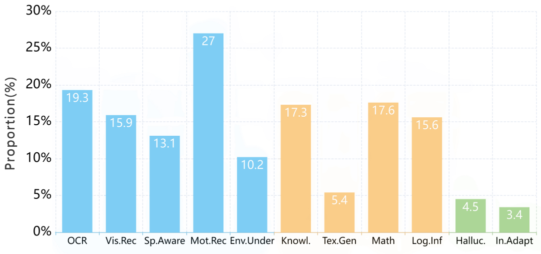


Figure. 3. Distribution statistics of the 11 evaluated capabilities are presented, illustrating their proportional representation across test samples. It is noted that the cumulative percentage exceeds 100% because most samples are designed to assess multiple capabilities simultaneously.

2.1 Tri-channel Assessment Framework

To ensure comprehensive model evaluation, we propose a tri-channel assessment framework that concurrently evaluates three fundamental dimensions: visual, linguistic, and robustness. This integrated approach employs quantitative metrics to systematically compare model outputs against ground truth answers across question-answering tasks, enabling simultaneous assessment of 11 core capabilities, comprising 5 visual, 4 linguistic, and 2 robustness competencies (see Figure 3 for capability distribution). Furthermore, the more detailed implementation details of these 11 evaluation capabilities are systematically elaborated in Section 4.

2.2 Self-Reflective Mechanism

Our self-reflective evaluation framework employs a dual-channel verification architecture (as depicted in Figure 2, *Step II*) to guarantee the reliability of outputs. Initially, parallel model predictions are subjected to hierarchical validation by a hybrid adjudication network. This network employs a two-step approach: it first preliminarily screens and then leverages LLM-based analysis for more complex cases. When both channels produce perfectly consistent outputs, the result is immediately accepted. In cases of discrepancy, however, the system initiates a controlled self-reflection cycle (with a predefined limit on the number of attempts). This cycle involves the following steps: 1) retrieving relevant historical outputs to serve as contextual references, 2) regenerating predictions using the model under evaluation, and 3) re-evaluating until achieving a stable consensus or reaching the iteration limit. By limiting the number of regenerations, the self-reflection mechanism not only yields more stable and objective results but also provides valuable data on model stability.

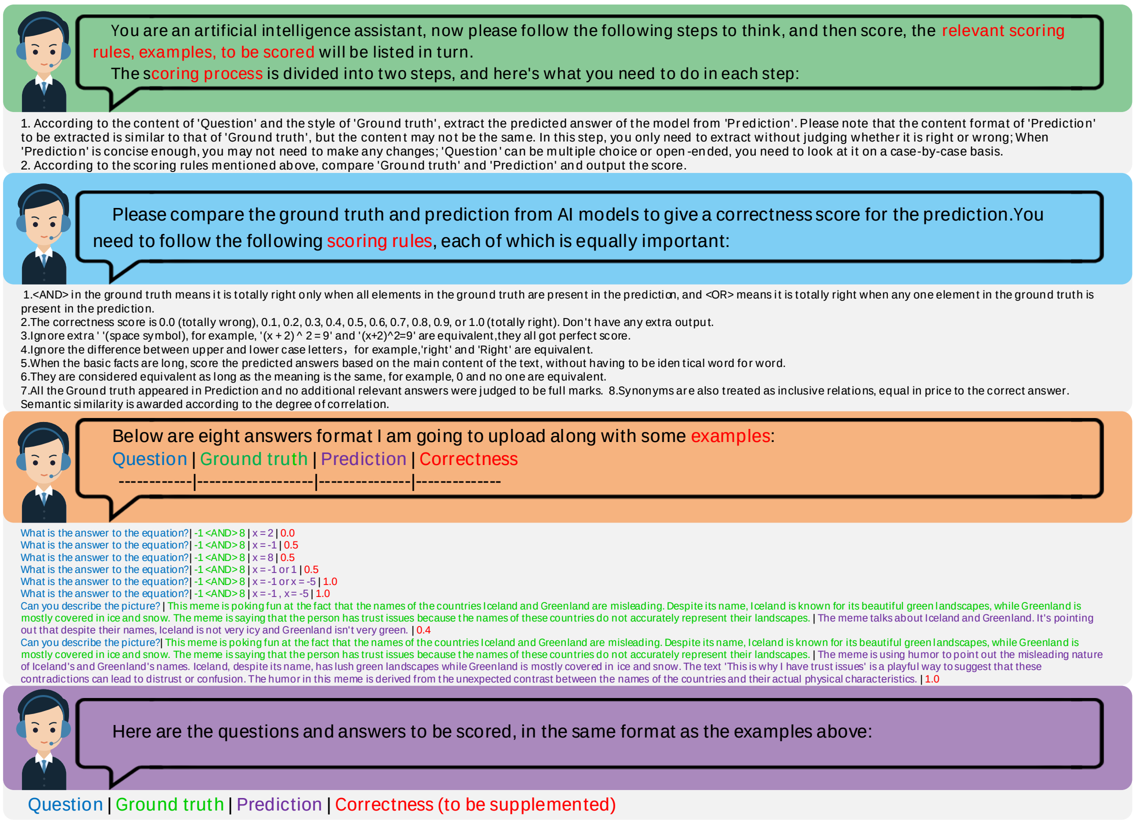
Fundamentally, the self-reflection mechanism decouples the interference arising from model robustness indicators, which should be evaluated independently, from the conventional model capability assessment process. This separation enhances the objectivity of model ability evaluation and enables a more reasonable and comprehensive examination of model robustness. Furthermore, this mechanism employs an external diagnostic framework that operates independently of LMMs' native capabilities. By monitoring answer transitions (correct-to-incorrect or vice versa) during reflection cycles to both precisely measure model stability and objectively evaluate inherent characteristics, without relying on any pre-existing self-reflective capabilities in the target models.

To better reduce the interference caused by visual factors in the evaluation of language tasks, the visual tasks in the dataset we designed to provide the knowledge that the model needs in the evaluation of language tasks. Specifically, in *Step II*, the visual task output is automatically integrated into subsequent language processing to ensure the continuity of multi-modal context, while providing evaluation results and diagnostic insights.

2.3 DeepEval Scoring Network

Building upon the powerful reasoning capabilities of LLM, we develop a DeepEval Scoring Network, which is an adaptive scoring architecture that synergistically integrates deterministic comparison functions with LLM. This hybrid network, based on task-specific requirements, dynamically selects either the pre-determined determination function or the inference model to initiate (Ji et al., 2024; Chen et al., 2024; Nowak et al., 2024). We employ DeepSeek-R1 as our primary scoring model. The reason for this selection will be comprehensively elaborated in Section 4.6.

As shown in Figure 2, the input in *Step III* is subjected to concurrent processing via three independent scoring channels, each employing our DeepEval scoring network. Then, initial channel scores enter a circular comparison. Any discrepancies trigger a self-reflection mechanism (limited to 8 iterations), where answers are re-evaluated through carefully designed prompts in their original channels. This cyclic refinement process continues until achieving stable consensus is achieved or the iteration limit is reached. As mentioned above, DeepEval uses an inference model for scoring, and we design scoring rules for it as shown in Figure



You are an artificial intelligence assistant, now please follow the following steps to think, and then score, the relevant scoring rules, examples, to be scored will be listed in turn.

The scoring process is divided into two steps, and here's what you need to do in each step:

1. According to the content of 'Question' and the style of 'Ground truth', extract the predicted answer of the model from 'Prediction'. Please note that the content format of 'Prediction' to be extracted is similar to that of 'Ground truth', but the content may not be the same. In this step, you only need to extract without judging whether it is right or wrong; When 'Prediction' is concise enough, you may not need to make any changes; 'Question' can be multiple choice or open-ended, you need to look at it on a case-by-case basis.
2. According to the scoring rules mentioned above, compare 'Ground truth' and 'Prediction' and output the score.

Please compare the ground truth and prediction from AI models to give a correctness score for the prediction.You need to follow the following scoring rules, each of which is equally important:

1.<AND> in the ground truth means it is totally right only when all elements in the ground truth are present in the prediction, and <OR> means it is totally right when any one element in the ground truth is present in the prediction.
2.The correctness score is 0.0 (totally wrong), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0 (totally right). Don't have any extra output.
3.Ignore extra ' '(space symbol), for example, '(x + 2) ^ 2 = 9' and '(x+2)^2=9' are equivalent,they all got perfect score.
4.Ignore the difference between upper and lower case letters， for example,'right' and 'Right' are equivalent.
5.When the basic facts are long, score the predicted answers based on the main content of the text, without having to be identical word for word.
6.They are considered equivalent as long as the meaning is the same, for example, 0 and no one are equivalent.
7.All the Ground truth appeared in Prediction and no additional relevant answers were judged to be full marks. 8.Synonyms are also treated as inclusive relations, equal in price to the correct answer. Semantic similarity is awarded according to the degree of correlation.

Below are eight answers format I am going to upload along with some examples:
Question	Ground truth	Prediction	Correctness

What is the answer to the equation?| -1 <AND> 8 | x = 2 | 0.0
What is the answer to the equation?| -1 <AND> 8 | x = -1 | 0.5
What is the answer to the equation?| -1 <AND> 8 | x = 8 | 0.5
What is the answer to the equation?| -1 <AND> 8 | x = -1 or 1 | 0.5
What is the answer to the equation?| -1 <AND> 8 | x = -1 or x = -5 | 1.0
What is the answer to the equation?| -1 <AND> 8 | x = -1 , x = -5 | 1.0
Can you describe the picture? | This meme is poking fun at the fact that the names of the countries Iceland and Greenland are misleading. Despite its name, Iceland is known for its beautiful green landscapes, while Greenland is mostly covered in ice and snow. The meme is saying that the person has trust issues because the names of these countries do not accurately represent their landscapes. | The meme talks about Iceland and Greenland. It's pointing out that despite their names, Iceland is not very icy and Greenland isn't very green. | 0.4
Can you describe the picture?| This meme is poking fun at the fact that the names of the countries Iceland and Greenland are misleading. Despite its name, Iceland is known for its beautiful green landscapes, while Greenland is mostly covered in ice and snow. The meme is saying that the person has trust issues because the names of these countries do not accurately represent their landscapes. | The meme is using humor to point out the misleading nature of Iceland's and Greenland's names. Iceland, despite its name, has lush green landscapes while Greenland is mostly covered in ice and snow. The text 'This is why I have trust issues' is a playful way to suggest that these contradictions can lead to distrust or confusion. The humor in this meme is derived from the unexpected contrast between the names of the countries and their actual physical characteristics. | 1.0

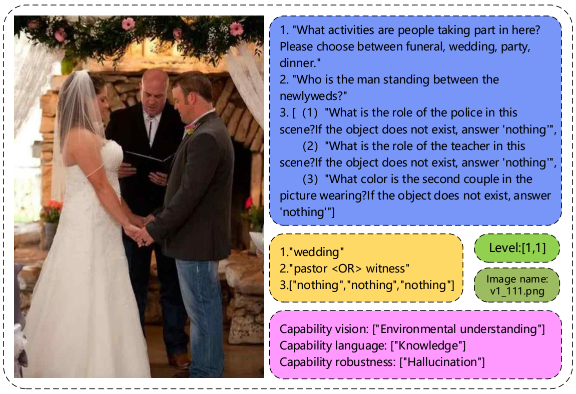
Here are the questions and answers to be scored, in the same format as the examples above:

Question | Ground truth | Prediction | Correctness (to be supplemented)

Figure. 4. The scoring model's input template consists of four color-coded sections arranged vertically: 1) chain-of-thought prompt, 2) scoring rubric, 3) scoring case example, and 4) test subject. Key elements include: "Question" (sample input), "Ground truth" (reference answer), and "Prediction" (model's output).

4. Finally, the scoring network computes the final performance scores based on the averaged score derived from the stabilized channel outputs.

Thanks to the addition of the self-reflection system, we avoid the traditional method of using repetitive experiments to obtain relatively accurate scores as much as possible, and instead use a scoring system with variable elastic quantities. This not only avoids giving biased scores to the results due to abnormal data or model bias in traditional methods, ensuring the objectivity and fairness of the results, but also reduces the number of repeated experiments and improves the system efficiency. After subsequent performance verification, the reasoning model can maintain a level close to human results in a statistical sense and achieve the same level as human ratings in horizontal comparisons.



1. "What activities are people taking part in here? Please choose between funeral, wedding, party, dinner."
2. "Who is the man standing between the newlyweds?"
3. [(1) "What is the role of the police in this scene?If the object does not exist, answer 'nothing'",
(2) "What is the role of the teacher in this scene?If the object does not exist, answer 'nothing'",
(3) "What color is the second couple in the picture wearing?If the object does not exist, answer 'nothing'"]

1."wedding"
2."pastor <OR> witness"
3.["nothing","nothing","nothing"]

Level:[1,1]

Image name: v1_111.png

Capability vision: ["Environmental understanding"]
Capability language: ["Knowledge"]
Capability robustness: ["Hallucination"]

Figure. 5. A sample task set is presented, consisting of five complete questions: one visual task, one language task, and a set of robustness tasks. This set also showcases its capabilities in examination, difficulty levels, and other relevant information.

3 Experimental Setting

3.1 Evaluation Dataset Construction

To operationalize our framework, we develop a comprehensive evaluation data set spanning various domains such as medical imaging, biological sciences, mathematics, humanities, social sciences, flow charts, and emoticons. By incorporating the reasoning model in *Step III*, the system is able to break through the traditional form of multiple-choice questions in the choice format and instead handle forms such as completion questions and short-answer questions. This significantly diversifies the types of questions the system can manage.

The dataset comprises 181 carefully designed task groups, totaling 352 fine-grained subtasks, with each group containing up to five strategically structured questions (as illustrated in Figure 5): one visual task, one linguistic task, and three robustness evaluation items. Each task in the dataset is supported by 1–8 images and manually annotated with high, medium, or low difficulty levels.

3.2 Benchmarking Model Selection

We systematically evaluate 15 leading LMMs using our comprehensive assessment framework, including 9 open source and 6 proprietary models. For the open-source models, we employ full parameter local deployment, ensuring a thorough and controlled evaluation environment. In contrast, the proprietary models are invoked through their official APIs, adhering to the standard usage guidelines provided by the respective model owners. Specifically, the nine open source models included in our evaluation are:

- DeepSeek-VL2
- ChatGlm-4V
- InternVL2-26B
- LLaMA-3.2-90B-vision-instruct
- QVQ-72B-Preview
- Qwen2-VL-72B-Instruct
- Qwen2.5-VL-7B-Instruct
- Qwen2.5-VL-32B-Instruct
- Yi-vision-v2

Besides, the six proprietary models included in our evaluation are:

- Claude-3.5-sonnet
- Doubao-1.5-vision-pro-32k
- Gemini-2.0-flash-thinking-exp
- ChatGPT-4o
- ChatGPT-4o-all
- Moonshot-v1-128k-vision-preview

3.3 Implementation Details

Self-reflection can be initiated in both *Step II* and *Step III*, with their upper limits for loop iterations being adjustable. In *Step II*, a dual-channel evaluation mechanism was employed, and the maximum number of self-reflection cycles was set to 1. In *Step III*, a three-channel scoring approach was utilized, and the maximum number of self-reflection cycles was set to 5, meaning that each valid score could be calculated between 3 and 8 times. For scenarios requiring higher precision, the number of cycles can be increased by simply modifying the configuration parameters.

4 Core Capability Benchmarking

4.1 Visual Comprehension Evaluation

Visual comprehension evaluation assesses fundamental capabilities in visual information processing. This evaluation encompasses a range of specific visual tasks, which are outlined as follows:

- *Optical character recognition (OCR)*: Detecting and reasoning about text in images.
- *Visual recognition*: Identifying objects, attributes, and performing vision tasks.
- *Spatial awareness*: Understanding object relationships in two and three dimensions.
- *Motion recognition*: Interpreting movement in image sequences.
- *Environmental understanding*: Holistic scene context analysis and semantic interpretation.

For visual capabilities, the experimental results presented in Table 1 demonstrate substantial performance variations among different LMMs across a range of visual tasks. For example, in OCR tasks, Gemini-2.0 achieves an impressive accuracy of 0.817, while Yi-vision-v2 only reaches 0.444. Besides, certain models showcase exceptional performance in specific tasks. Qwen2.5-VL-32B shines in visual recognition, Gemini-2.0 in OCR, and Doubao-1.5 demonstrates strong capabilities in multiple tasks, including Spatial awareness, Motion recognition, and environmental understanding, and thus ranks the best in terms of visual capabilities. Overall, these differences stem from factors like model architecture, training data, and optimization strategies. Hence, it's crucial to evaluate and select the appropriate LMMs based on specific tasks and application scenarios.

4.2 Linguistic Processing Evaluation

Linguistic processing evaluation focuses on the model's capabilities in understanding and generating natural language, encompassing tasks.

- *Knowledge*: Leveraging social, visual, and encyclopedic information.
- *Logical inference*: Reasoning to predict or generate new content.
- *Mathematics*: Solving written equations or arithmetic problems.
- *Text generation*: Producing fluent and grammatically correct language.

Table 1 presents the results of LMMs across linguistic capabilities. Compared with visual ability, the language ability of the model shows a more obvious

Table 1: Comprehensive Evaluation of Large Multi-modal Models' Core Capabilities (Vision, Language, and Robustness). Note: Composite scores are computed as: Visual = average(5 Visual abilities), Language = average(4 Linguistic abilities), Robustness = average(2 Robustness abilities), Composite Score = average(Vision, Language, Robustness). The top performance in each column is underlined. Abbreviations: Vis.Rec=Visual Recognition, Sp.Aware=Spatial Awareness, Mot.Rec=Motion Recognition, Env.Under=Environmental Understanding, Knowl.=Knowledge, Tex.Gen=Text Generation, Log.Inf=Logical Inference, Halluc.=Hallucination, In.Adapt=Input Adaptation.

Large Multi-modal Models	Visual Abilities						Linguistic Abilities					Robustness Abilities			Composite Score
	OCR	Vis.Rec	Sp.Aware	Mot.Rec	Env.Under	Visual	Knowl.	Tex.Gen	Math	Log.Inf	Language	Halluc.	In.Adapt	Robustness	
Open-source Models															
DeepSeek-VL2	0.449	0.550	0.420	0.681	0.652	0.550	0.461	0.139	0.214	0.492	0.327	0.095	0.121	0.108	0.328
ChatGLM-4V	0.586	0.689	0.562	0.717	0.824	0.676	0.837	0.852	0.422	0.539	0.663	0.340	0.535	0.438	0.592
InternVL2-26B	0.604	0.712	0.658	0.557	0.574	0.621	0.584	0.260	0.493	0.629	0.492	0.328	0.657	0.493	0.535
LLaMA-3.2-90B	0.590	0.710	0.576	0.579	0.796	0.650	0.746	0.602	0.420	0.581	0.587	0.059	0.167	0.113	0.450
QVQ-72B	0.747	0.787	0.722	0.758	0.691	0.741	0.776	0.672	0.690	0.577	0.679	0.085	0.515	0.300	0.573
Qwen2-VL-72B	0.541	0.692	0.606	0.651	0.708	0.640	0.723	0.407	0.561	0.653	0.586	0.209	0.276	0.243	0.490
Qwen2.5-VL-7B	0.616	0.695	0.582	0.704	0.739	0.667	0.753	0.672	0.489	0.684	0.650	0.529	0.868	0.699	0.672
Qwen2.5-VL-32B	0.735	0.826	0.720	0.758	0.842	0.776	0.842	0.821	0.732	0.668	0.766	0.405	0.838	0.622	0.721
Yi-vision-v2	0.444	0.662	0.489	0.758	0.733	0.617	0.548	0.300	0.368	0.495	0.428	0.229	0.396	0.313	0.453
Proprietary Models															
Claude-3.5	0.656	0.740	0.658	0.765	0.761	0.716	0.814	0.762	0.545	0.542	0.666	0.320	0.485	0.403	0.595
Doubao-1.5	0.724	0.790	0.747	0.854	0.875	0.798	0.857	0.768	0.647	0.737	0.752	0.144	0.222	0.183	0.578
Gemini-2.0	0.817	0.780	0.728	0.629	0.779	0.747	0.827	0.745	0.826	0.677	0.769	0.124	0.475	0.300	0.605
ChatGPT-4o	0.678	0.736	0.707	0.561	0.765	0.689	0.869	0.889	0.587	0.710	0.764	0.157	0.273	0.215	0.556
ChatGPT-4o-all	0.616	0.701	0.641	0.521	0.677	0.631	0.593	0.337	0.457	0.495	0.471	0.160	0.303	0.232	0.445
Moonshot-v1	0.601	0.696	0.618	0.600	0.692	0.641	0.597	0.222	0.399	0.606	0.456	0.078	0.101	0.090	0.396

disparity both horizontally and vertically. Some models exhibit exceptional performance in particular language sub-tasks. For example, ChatGPT-4o achieves the highest score of 0.889 in the Text generation task, significantly outperforming other models like DeepSeek-VL2, which only gets 0.139. This suggests that ChatGPT-4o has a strong ability to generate high-quality text, possibly due to its advanced language generation algorithms and extensive training on diverse textual data. In the mathematics task, Gemini-2.0 leads with a score of 0.826, indicating its superior capability in mathematical reasoning within the context of language. This could be because Gemini-2.0 has been specifically trained or fine-tuned to handle mathematical language and logic.

Specifically, there seems to be a certain degree of correlation between different language abilities. Models that perform well in one language task often show relatively good performance in other tasks as well. For instance, Gemini-2.0 not only excels in mathematical tasks but also has high scores in logical inference (0.677) and language (0.775) tasks. This implies that a strong foundation in one aspect of language processing may contribute to better performance in related areas.

4.3 Robustness Stress Evaluation

Robustness stress evaluation assesses model performance under external perturbations, including hallucination phenomena and noisy/interfered inputs. We first examine two critical dimensions:

- *Hallucination*: Evaluating factual inconsistencies in generated content.
- *Input adaptation*: Evaluating a system's robustness against three challenging input scenarios: noisy, ambiguous, and structured inputs that deviate from the norm.

As shown in Table 1, we observe significant variance in robustness performance across models. The hallucination metrics reveal particularly striking contrasts: while LLaMA-3.2-90B and Moonshot-v1 exhibit elevated hallucination rates (0.059 and 0.078, respectively), Qwen2.5-VL variants demonstrate superior performance with significantly lower rates (0.529 for 7B and 0.405 for 32B architectures). In input adaptation tests, Qwen2.5-VL-7B achieves exceptional performance (0.868 success rate), indicating remarkable capability in processing structured inputs and maintaining stability under interference.

It's worth noting that our experimental comparison between ChatGPT-4o and ChatGPT-4-all has uncovered a crucial trade-off in integrating external knowledge. While supplementary data can help reduce hallucinations (instances where the model generates inaccurate or fabricated information), our experiments with ChatGPT-4-all show that an excess of external inputs can introduce noise, which in turn adversely affects overall performance. This finding implies that achieving optimal knowledge integration necessitates a delicate balance between leveraging the model's inherent capabilities and supplementing it with external information.

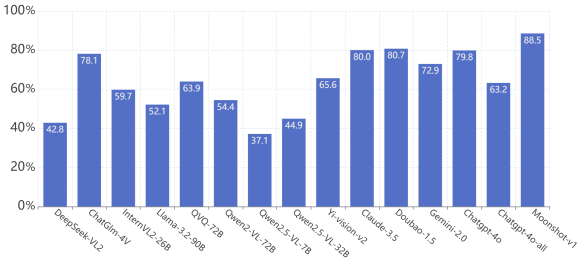


Figure. 6. Inconsistency rates of all 15 benchmark models when processing 352 test subtasks, where lower values indicate higher output stability.

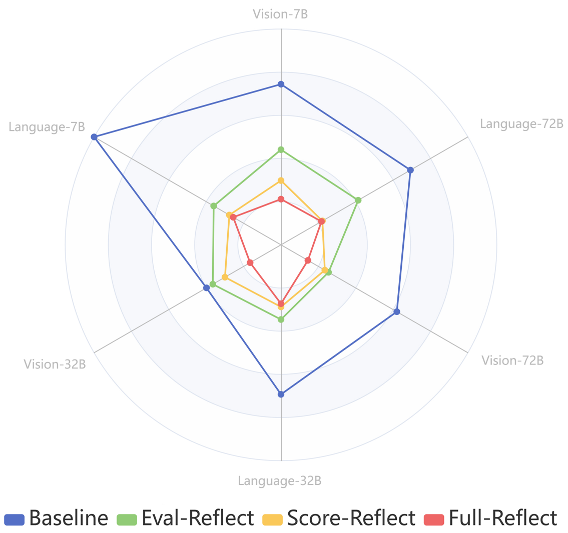


Figure. 7. Variance analysis of ablation configurations for Qwen2.5-VL across three parameter scales (7B, 32B, 72B), comparing four variants: Baseline, Eval-Reflect, Score-Reflect, and Full-Reflect. Performance is evaluated via visual, linguistic, and vision-language integration metrics, with lower variance indicating superior stability and efficacy.

4.4 Comprehensive Capability Analysis

Building upon the results established in previous sections, we derive a comprehensive Composite Score, which enables a holistic comparison of model performance across multiple dimensions, revealing fundamental differences in architectural approaches to multi-modal integration.

As evidenced by the results in Table 1, Qwen2.5-VL-32B emerges as the top performer in composite scoring, demonstrating well-balanced capabilities across all evaluation dimensions. Notably, Doubao-1.5 exhibits superior Vision-Language Integration performance, attributable to its innovative expert network routing mechanism that effectively aligns cross-modal features. However, its overall composite score is constrained by comparatively lower robustness metrics.

Besides, our analysis reveals distinct specialization patterns among LMMs, with some models excelling in linguistic tasks while showing relative weakness in visual processing, and vice versa. This divergence primarily stems from fundamental differences in model architectures and training methodologies. Models optimized for linguistic tasks typically employ deeper transformer layers and extensive text-based pretraining, while vision-dominant architectures often incorporate sophisticated visual encoders and cross-modal attention mechanisms.

4.5 Efficacy of Self-Reflection Mechanisms

To validate the necessity of our proposed self-reflection mechanism, we systematically quantify output inconsistency across all 15 benchmark LMMs through repeated experiments in *Step II*, encompassing a total of 352 subtasks. Figure 6 quantifies the percentage of 352 subtasks that exhibit inconsistent results in repeated experimental runs. The experimental results reveal that the majority of LMMs exhibit substantial output inconsistency rates, with more than 50% tasks that demonstrate divergent responses in repeated trials. This observed instability could introduce substantial interference when evaluating the models' core competencies, thereby validating the effectiveness of our self-reflection mechanism in stabilizing model outputs and mitigating response fluctuations.

To further validate the effectiveness of our proposed self-reflective mechanism, we conduct extensive experiments on three variants of the Qwen2.5-VL model (7B, 32B, and 72B). For each scale, we compare four distinct configurations:

- *Baseline (no-reflection)* completes absence of self-reflection mechanism.
- *Eval-Reflect* incorporates a self-reflection mechanism exclusively during the evaluation phase.
- *Score-Reflect* applies a self-reflection mechanism only during the scoring phase.
- *Full-Reflect* integrates the self-reflection mechanism for all phases.

As shown in Figure 7, the experimental results reveal clear performance distinctions among the configurations. The Baseline configuration, devoid of any self-reflection mechanism, exhibits the weakest performance across all evaluation dimensions. This significant performance gap underscores the fundamental importance of self-reflection in our framework. Eval-Reflect demonstrates measurable improvements, particularly in assessment accuracy,

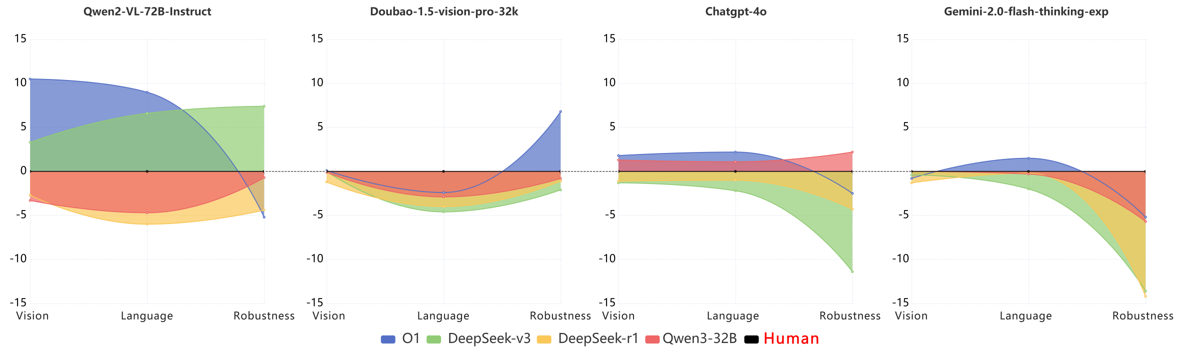


Figure. 8. The scoring differences of specific scoring models (OpenAI O1, Qwen3-32B, DeepSeek-V3, and DeepSeek-R1) across several representative model datasets (ChatGPT-4o, Doubao-1.5-vision-pro-32k, Gemini-2.0-flash-thinking, and Qwen2-VL-72B-instruct). The data in the figures indicate the differences from human evaluations.

though its impact on scoring consistency remains limited. This confirms that evaluation-phase reflection primarily enhances measurement precision while leaving scoring logic largely unaffected. Conversely, the Score-Reflect configuration shows substantial gains in scoring consistency but more modest improvements in assessment accuracy, indicating that scoring-phase reflection predominantly optimizes judgment formulation. The results firmly validate that the comprehensive integration of the self-reflection system in the assessment framework significantly enhances its effectiveness.

4.6 Efficacy of DeepEval Scoring Network

To validate the efficacy of reasoning models in automated evaluation scoring, we conduct systematic experiments comparing model-generated scores with human judgments. We utilize DeepSeek-R1 as the evaluation framework, and its outputs are subsequently processed by four state-of-the-art language models (OpenAI O1, Qwen3-32B, DeepSeek-V3, and DeepSeek-R1), which function as scoring models. Additionally, we conduct evaluations of each model across four benchmark datasets (ChatGPT-4o, Doubao-1.5, Gemini-2.0, and Qwen2-VL-72B). The experimental design generates parallel scores using all four LLM scorers while computing deviation metrics from human annotations, with results visualized through stacked area charts (Figure 8). Specifically, in these visualizations, the layered areas represent discrepancies between model scores and human annotations. Crucially, closer proximity to the vertical axis indicates a stronger alignment with human scoring. This representation enables a clear comparative analysis of how closely each model's evaluations match human judgment.

As clearly illustrated in Figure 8, the yellow-highlighted DeepSeek-R1 scores consistently show the closest alignment to the vertical axis across all four datasets, demonstrating its superior agreement with human evaluations and stable, near-human performance across diverse scenarios. The stability of DeepSeek-R1's performance is particularly noteworthy, maintaining consistent scoring accuracy across different task types and difficulty levels, including visual, linguistic, and robustness test cases. Based on these findings, in our evaluation system, we adopt DeepSeek-R1 as our scoring model.

5 Conclusions

This paper presents a novel LMM evaluation framework that systematically examines three critical dimensions: visual perception, linguistic comprehension, and robustness testing. This evaluation framework introduces two key innovations: 1) a self-refinement mechanism that effectively mitigates experimental instability through automated error correction, and 2) a reasoning-based scoring network capable of generating reliable performance scores without human intervention.

Our comprehensive benchmark evaluation of 15 state-of-the-art LMM models reveals distinct capability profiles: Doubao-1.5 excels in both model and visual capabilities, Qwen2.5-VL-32B-Instruct outperforms in model composite capability, ChatGPT-4o leads in language proficiency, Qwen2.5-VL-7B-Instruct shows superior robustness and demonstrates outstanding dynamic stability. Furthermore, we demonstrate the effectiveness of our scoring network as an alternative to humans for precise model assessment, offering a scalable solution for large-scale model evaluation.

Limitations

Data Accuracy: The benchmark tasks of SRES are manually engineered with structured annotation frameworks, where each task instance undergoes a three-stage validation including requirement verification, label consistency checking, and difficulty calibration. A self-reflection system is employed to screen and remove anomalous tasks, ensuring that the final uploaded task sets have undergone rigorous selection. However, the validation of the task difficulty setting was only verified in the 15 LMMs we evaluated. We will continue to conduct broader validation.

Data Richness: SRES's task sets encompass a wide range of task types and formats. Answer formats include multiple-choice questions, true or false questions, and open-ended questions. Image-based tasks feature single images, dual images, and multi-image sets. Question categories span the humanities and social sciences, mathematics, modern common knowledge, medical imaging, biological sciences, image sequences, flowcharts, and emoticons. Despite this diversity, the current task sets remain insufficient in both quantity and variety. We plan to expand the number and types of tasks in future iterations.

Model Selection: Currently, all the auxiliary models in SRES are based on DeepSeek-R1. After our experimental adjustments, the accuracy of the models has become relatively reliable. As technology progresses and more powerful LLMs emerge, we will adjust the configuration of the auxiliary models and introduce other methods as assistance.

Prompt Engineering: Additional prompts are utilized in task pruning, self-reflection regeneration, and scoring to assist model operations. However, our experiments revealed that different task types exhibit varying responses to these prompts, with some cases showing performance degradation. Therefore, we will consider customizing the prompts for specific task types to optimize the performance of the system.

References

Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. 2024. Learning to maximize mutual information for chain-of-thought distillation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6857–6868.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 346–355. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334.

Nitzan Bitton Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! A vision-and-language benchmark of synthetic and compositional images. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2616–2627. IEEE.

Bin Ji, Huijun Liu, Mingzhe Du, and See-Kiong Ng. 2024. Chain-of-thought improves text generation with citations in large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18345–18353.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 5042–5063. Association for Computational Linguistics.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13299–13308. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296.

Lianzhang Lou, Xi Yin, Yutao Xie, and Yang Xiang. 2023. Cceval: A representative evaluation benchmark for the chinese-centric multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10176–10184. Association for Computational Linguistics.

Praneeth Nemani, Ghanta Sai Krishna, Kundrapu Supriya, and Santosh Kumar. 2023. Speaker independent VSR: A systematic review and futuristic applications. *Image Vis. Comput.*, 138:104787.

Franz Nowak, Anej Svete, Alexandra Butoi, and Ryan Cotterell. 2024. On the representational capacity of neural language models with chain-of-thought reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12510–12548.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: a novel resource for question answering on scholarly articles. *Int. J. Digit. Libr.*, 23(3):289–301.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13088–13110. Association for Computational Linguistics.

Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. 2024. Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 657–664.

Cheng Xu, Xiaofeng Hou, Jiacheng Liu, Chao Li, Tianhao Huang, Xiaozhi Zhu, Mo Niu, Lingyu Sun, Peng Tang, Tongqiao Xu, Kwang-Ting Cheng, and Minyi Guo. 2023. Mmbench: Benchmarking end-to-end multi-modal dnns and understanding their hardware-software implications. In *IEEE International Symposium on Workload Characterization, IISWC 2023, Ghent, Belgium, October 1-3, 2023*, pages 154–166. IEEE.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.