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

Anonymous ACL submission

Abstract

This paper introduces SSR-VLES, a structured multi-perspective and multi-modal comprehensive evaluation system based on self-reflection, designed to assess the overall capabilities of large multi-modal models (LMMs) in complex multi-modal tasks. SSR-VLES addresses this gap by defining 11 composite tasks that encompass five visual functions, four language functions and robustness, while also model dynamic stability. The system evaluates LMMs across four dimensions: visual ability, language ability, robustness and model dynamic stability. It employs a self-reflection mechanism to ensure stable model outputs and enhances evaluation accuracy and flexibility through multi-round dialogue mechanisms and additional prompts. Experimental results demonstrate that SSR-VLES can effectively differentiate the capability levels of various LMMs and provide valuable guidance for further model optimization. SSR-VLES code are available at https://anonymous.4open.science/r/SSR-VLES-BF91

1 Introduction

Large Multi-modal Models (LMMs) have made remarkable progress in recent years, with numerous models being proposed to demonstrate their effectiveness from diverse perspectives (Dai et al., 2023; Zhu et al., 2024; Li et al., 2023a). Despite this progress, there is a significant lack of a comprehensive evaluation system that accurately quantifies the performance of these LMMs (Liu et al., 2024a; Yu et al., 2024; Liu et al., 2024b; Schwenk et al., 2022).

However, current evaluation systems mainly concentrate on single-modal tasks, such as image or text analysis, while neglecting the necessity of comprehensive multi-modal task assessment. The limitations can be further elaborated upon in terms of both breadth and depth. 1) **Horizontal Dimension (Task Breadth)**: Current systems predominantly focus on a narrow range of modal combinations, primarily text-image pairs. This narrow focus means that the vast majority of practical multi-modal application scenarios are left unexplored. 2) **Vertical Dimension (Interaction Depth)**: Multi-modal tasks vary significantly, necessitating tailored evaluation criteria. Current systems often apply generic metrics that may not fully capture the nuances of individual tasks. Moreover, complex multi-modal tasks, which involve interactions across multiple modalities, require a balanced approach that considers multiple dimensions simultaneously.

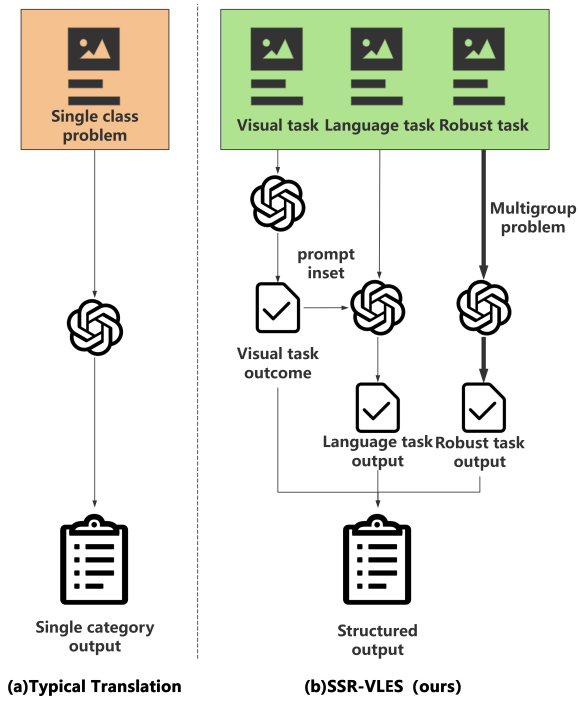

Figure. 1. The current mainstream evaluation system is picture (a), and the SSR-VLES evaluation system is picture (b). "task" refers to a single problem in the input model.

Given these limitations, there is an urgent need for a detailed and comprehensive evaluation framework that addresses both the breadth and depth of

multi-modal evaluation. Therefore, in this paper, we propose a novel benchmark framework, SSR-VLES (*Structured Self-Reflective Vision-Language Evaluation System*), to provide a comprehensive assessment of the overall capabilities of LLMs.

The following elaborates on the key aspects of SSR-VLES.

- We innovatively design a self-reflection evaluation mechanism. This mechanism establishes a dynamic feedback correction system to effectively mitigate the interference of model output fluctuations on evaluation results, enhancing the accuracy, objectivity, and reliability of the evaluation system.
- We provide a more comprehensive and realistic evaluation of model performance by defining a hierarchical evaluation architecture comprising three core innovation modules: visual processing (5 visual capability dimensions), linguistic understanding (4 linguistic capability dimensions), and multi-modal interaction (2 anti-interference test scenarios and dynamic stability indices).
- We design an automated model evaluation system based on LLM, evaluate 13 major LMMs, fully analyzes the experimental results, and validate the system based on the results.

2 SSR-VLES

2.1 Structured Evaluation Framework

The architecture of LMMs typically integrates a visual translator alongside the core LLM(Large language model). This design inherently limits the model's visual capabilities to those of the visual translator, while its linguistic capabilities relies primarily on the LLM itself (Verma et al., 2024; Goyal et al., 2017). To ensure a nuanced evaluation of both the model's visual and linguistic strengths and weaknesses, we propose a structured evaluation framework that separately assesses four critical dimensions: visual processing, linguistic understanding, robustness, and dynamic stability.

Specifically, visual processing testing evaluates the model's ability to accurately interpret and process visual information, including tasks like object recognition, scene understanding, and image captioning, aiming to assess the effectiveness and limitations of the integrated visual translator within the LLM. The linguistic understanding testing, on the other hand, focuses on the model's capabilities in understanding and generating natural language, encompassing tasks such as language comprehension, text generation, sentiment analysis, and question answering, with the objective of gauging the core LLM's linguistic capabilities independently of its visual component. Robustness testing specifically targets potential weaknesses by presenting challenging scenarios(Li et al., 2023b). Model stability testing, on the other hand, focuses primarily on assessing the stability of the model, particularly the frequency of self-reflective systems, which is introduced in Section 2.2.

2.2 Self-Reflection Mechanism

The rationale behind introducing stability testing lies in the inherent limitations of current evaluation methodologies for black-box LMMs (Jiaming et al., 2024). A common practice among these methods is to average multiple results to achieve stability in evaluation outcomes. However, this traditional averaging technique often blends model instability with its core functional limitations, thus concealing the differences between these two distinct aspects. This blurring can lead to inaccurate evaluations of the model's true performance and capabilities.

To overcome this challenge, we introduce a *self-reflection mechanism*, that isolates stability assessment while improving result reliability and separately evaluates stability as part of the robustness dimension. As shown in Figure. 2(b), during the execution of a single atomic evaluation task, two independent yet identical evaluation channels are run simultaneously.The LLM-based referee then determines whether the two responses are equivalent.If they are not, the model under test is prompted to regenerate its output based on the previous result using a carefully designed prompt. Through a limited number of regenerations, the self-reflection mechanism achieves more stable and objective results while obtaining data on model stability. This approach avoids the high resource consumption and potential result distortion associated with traditional methods that rely on fixed multiple attempts. As shown in Figure. 2(c),a similar *self-reflection mechanism* has been used in the DeepEval-R1 Scoring Framework, which will be covered in Section 2.3.

2.3 DeepEval-R1 Scoring Framework

The diverse array of scenarios, which span both fixed-format responses and open-ended inquiries,

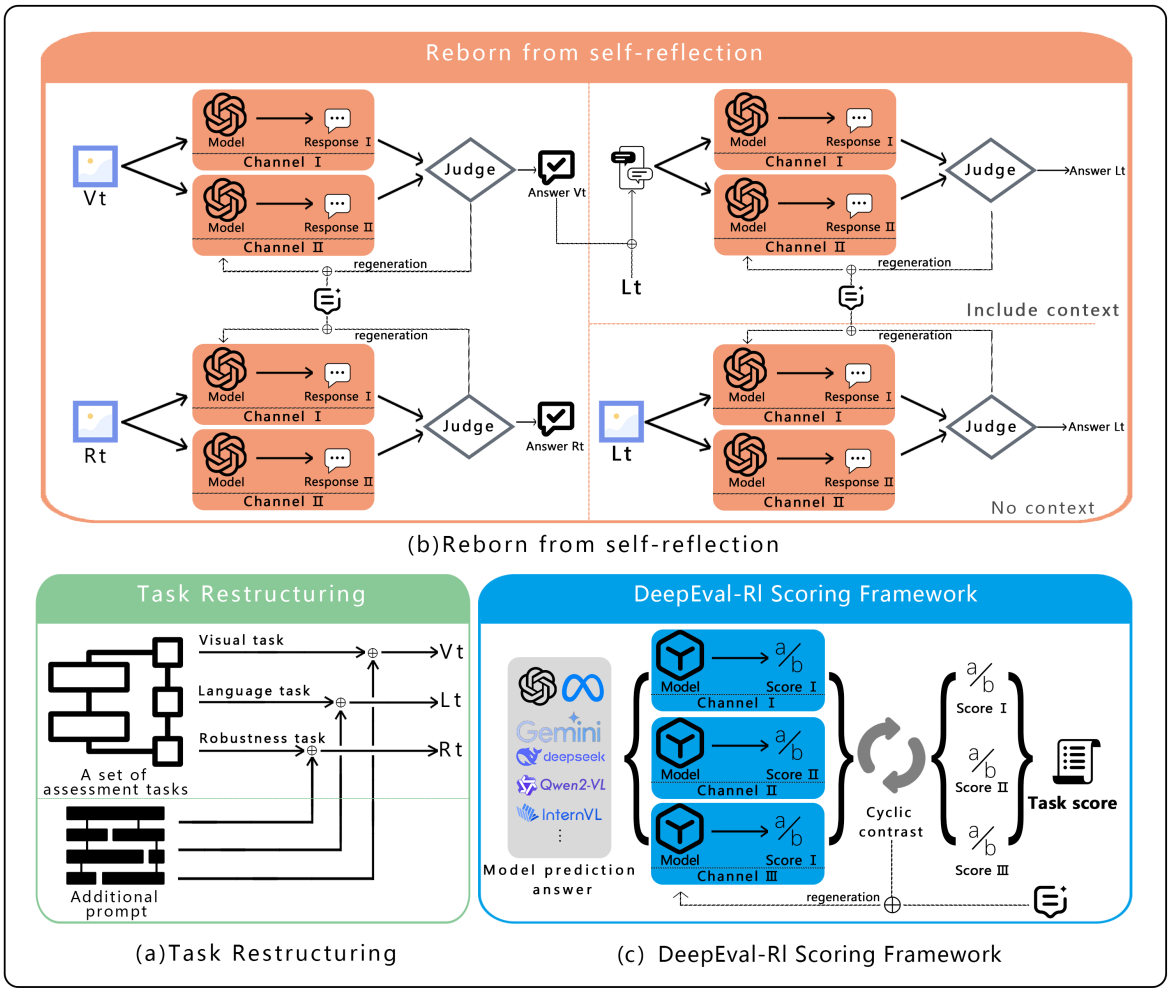

Figure. 2. (a) The structured task generation module constructs assessment tasks comprising three problem categories: visual (Vt), linguistic (Lt), and robustness (Rt). Each category is enhanced with targeted prompt engineering to create domain-specific challenges. (b) The self-reflective regeneration module processes these enhanced problems (Vt/Lt/Rt) to produce model predictions. This component enables iterative refinement of outputs through introspective reasoning mechanisms. (c) The tripartite evaluation framework employs parallel scoring channels, each combining a scoring model with an alternate verification model. This architecture computes performance metrics by comparing model predictions against ground truth values, and the results are optimized by the self-reflection mechanism.

presents significant challenges in model evaluation and metric design. Traditional methods are inadequate for accurately aligning the wide variety of predicted answers with the true answers, particularly given the complexity and nuances involved. Drawing inspiration from recent advancements in NLP and LMMs evaluation, we develop a sophisticated scoring framework based on DeepSeek-R1 (DeepSeek-AI et al., 2025; Dai et al., 2024) to enhance the evaluation process. DeepSeek-R1 has received widespread acclaim in recent academic circles, thanks to its innovative thought chain mechanism. This mechanism excels in achieving highly accurate interdisciplinary causal reasoning through a combination of hierarchical reasoning, multimodal correlation, and a dynamic calibration process (Ji et al., 2024; Chen et al., 2024a; Nowak et al., 2024).

To enhance the scalability of our scoring system, we meticulously design a composite prompt set tailored specifically for model evaluation. This prompt set carefully selects a variety of sample prompts, which are then fed through three distinct channels into the scoring model (DeepSeek-R1) to produce comprehensive scores. During the scoring process, the model initially checks for consistency among the scores generated by the three channels. This step is crucial for eliminating any erroneous ratings and ensuring the accuracy of the final score. In the event that the scores from the different channels differ, a review mechanism based on self-reflection mechanism is activated. This mechanism reconsiders the answers a limited number of times. The goal is to identify and correct

Figure. 3. The input template for the scoring model is divided into four parts, scoring rubric, chain of thought prompt,scoring case and test subject, from top to bottom, separated by color. Q represents the sample question;G represents the answer; P is the predicted value of the sample model.

any discrepancies, ensuring that the final score accurately reflects the model's performance. Finally, after all necessary reviews and adjustments have been made, the averaged score from the three channels is calculated. This averaged score serves as the definitive model's performance rating for the given question, providing a comprehensive and reliable assessment of the model's capabilities Alongside these sample prompts, as shown in Fig 3, we also establish a set of relevant scoring rules,chain of thought prompt and scoring case to ensure consistency and accuracy.

In addition, due to the uncontrollable nature of the model's output, the scoring model occasionally produces non-standardized outputs (Zhang et al., 2024). To address this, we design a compensation mechanism. When the output is non-standardized, the standby model (DeepSeek-v3) is activated to implement standardization procedures. If the model's output remains non-standardized, this mechanism judges the output and performs a limited number of retries. This will ensure that our evaluation system can handle automation in the face of non-standardized LLM output.

2.4 Overall Evaluation Process

The SSR-VLES framework's structured task testing process is designed to comprehensively evaluate LLMs rigorously and systematically. The following is an expanded and more detailed description of the overall evaluation process:

Step1. Task Restructuring: The initial input question for each of the three sub-tasks undergoes restructuring by appending an additional prompt tailored to the type of question to form a refined query. This newly crafted query is then submitted to the model under evaluation.

Step2. Parallel Task Channels: As shown in Figure. 2 (b), the query is processed simultaneously through two parallel task channels. Within these channels, the model generates predicted answers based on its internal processing mechanisms. In addition, when the set of problems includes both

visual and language tasks, the model output of the visual task is compiled as part of the input of the language task.

Step3. Output Comparison and Judgment: A judge is employed to meticulously compare the outputs from both channels. If the answers from both channels align perfectly, the model's output is deemed acceptable and is subsequently utilized. Conversely, if a discrepancy is observed, a self-reflection process is initiated. This process involves regenerating a limited number of answers until a reliable and consistent output is obtained.

Step4. Scoring and Evaluation: Once acceptable outputs are obtained, the deepEval-R1 scoring framework generates evaluation scores ranging from 0 to 1 based on predefined criteria. This framework leverages the advanced capabilities of DeepSeek-R1 to provide comprehensive and objective scores for each task.

3 Evaluation result

3.1 Experiments Settings

We use our evaluation system SSR-VLES to evaluate 13 mainstream LMMs: Claude3.5, deepseek-vl2 (Wu et al., 2024), Doubao1.5, Gemini2.0-flash (Sayyafzadeh et al., 2024), ChatGlm-4v, ChatGPT4o (OpenAI et al., 2024), ChatGPT4o-all, InternVL2 (Chen et al., 2024b), Llama-3.2, Moonshot-v1, QVQ, Qwen2-vl (Wang et al., 2024), and Yi-vision-v2.

We collect 110 images from diverse online sources and formulate 181 tasks (comprising a minimum of 318 sub-problems). Each task requires one or more specific capabilities to answer. These questions vary in type and complexity, necessitating open-ended or standard answers of different lengths. For each question, we identify the required capabilities and statistically summarize this information in Figure. 4. All true answers are manually annotated by experts. The question types encompass a wide range of categories, including humanities and social sciences, mathematics, modern common sense, medical imaging, biological science, image sequences, flowcharts, emoticons, and more, ensuring comprehensive coverage.

We develop 11 independent capability tests across three dimensions: visual ability, language ability, and robustness. For the visual task, we assess five core visual functions. These include visual recognition, which involves identifying objects, attributes, and performing advanced vision tasks; OCR, which focuses on recognizing and reasoning about text within images; spatial perception, understanding spatial relationships in both 2D and 3D contexts; motion recognition, identifying and interpreting movements in image sequences; and environmental understanding, recognizing and interpreting the contexts depicted in images. For the language task, we evaluate four core language functions. These encompass knowledge, utilizing social, visual, and encyclopedic information; inference, predicting or generating new content through reasoning; mathematics, solving written equations or arithmetic problems; and language generation, producing natural and correct language text. For the robustness task, we focus on two core robustness functions. These are hallucination, assessing when generated content is inconsistent with facts; and formatted input, evaluating robustness across varied input formats. Tasks are also classified by difficulty level: high (3), medium (2), and low (1).

In real-world scenarios, complex multi-modal tasks often require the integration of multiple core visual and language capabilities. Therefore, it is essential to include composite tasks that combine these capabilities in the evaluation framework. SSR-VLES designed 15 capability sets, as illustrated in Figure.5. Each set integrates multiple core capabilities, such as combining OCR with mathematical reasoning to solve icon problems; integrating visual recognition with knowledge to perform object tracking; and combining motion recognition with inference to predict future object movements. This approach allows for a more nuanced and comprehensive evaluation of LMMs.

Combining the aforementioned assessment tasks, we also report two comprehensive scores:

1) **Model capability**, which encompasses visual capability and language capability, provides a macro-level description of the LMMs' benchmark performance.

2) **Model composite score**, comprising visual capability, language capability, robustness, and dynamic stability indices, offers an all-encompassing evaluation of the model.

3.2 Multi-Perspective Evaluation

According to data in Table 1, the degree of synergy between visual and linguistic capabilities significantly impacts model performance. MoE architecture models demonstrate absolute superiority in cross-modal integration: Doubao1.5 ranks first in model capabilities, where its expert network

Model	Vision	Language	Model capability	Robustness	Model dynamic stability	Model composite score
Claude3.5	71.3%	65.8%	70.3%	38.5%	46.8%	68.0%
deepseek-vl2	53.3%	37.0%	47.9%	10.5%	17.0%	44.8%
Doubao1.5	78.7%	75.7%	76.6%	17.5%	30.1%	72.0%
Gemini2.0-flash	76.1%	77.5%	76.4%	26.2%	35.5%	72.3%
ChatGlm-4v	66.6%	64.1%	65.6%	41.7%	49.0%	64.0%
ChatGpt4o	70.3%	74.9%	73.1%	20.2%	32.2%	69.0%
ChatGpt4o-all	64.8%	50.2%	58.8%	21.6%	29.9%	56.0%
InternVL2	64.1%	53.9%	61.9%	45.7%	48.5%	60.6%
Llama-3.2	65.8%	59.6%	62.5%	10.1%	18.5%	58.1%
Moonshot-v1	65.0%	50.8%	59.0%	8.7%	24.7%	55.6%
QVQ	74.9%	68.5%	69.8%	25.4%	33.1%	66.1%
Qwen2-vl	64.2%	62.7%	61.8%	23.5%	29.7%	58.6%
Yi-vision-v2	60.4%	45.9%	53.8%	29.4%	36.7%	52.1%

Table 1: The multidimensional capabilities of the model to be tested, that is, visual capability, language capability, robustness, and model dynamic stability, are counted in 100%, and the highest score of a group of capabilities in the model to be tested is indicated by underline. The model ability is the integration of model vision ability and language ability.

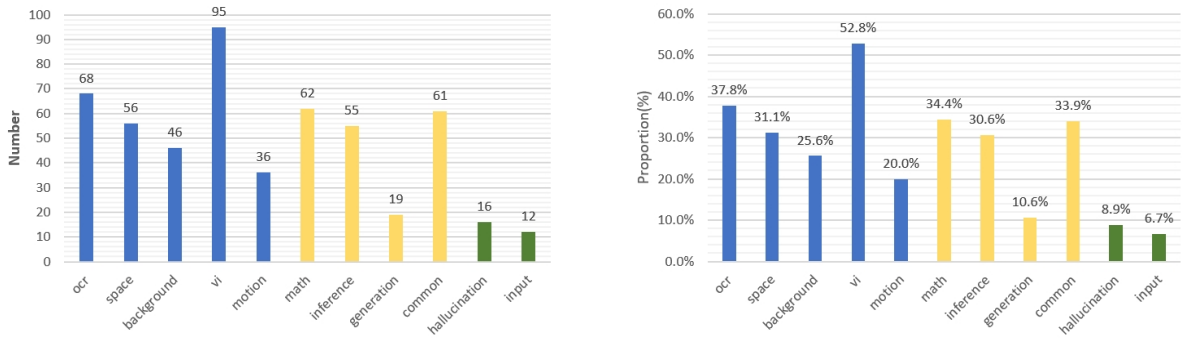

Figure. 4. The statistical distribution of our constructed 11 capability. (a) shows the frequency of each capability, while (b) illustrates the proportion of each capability. Note that the total percentage exceeds 100% because individual tasks may involve multiple labels.

routing mechanism effectively coordinates visual-linguistic feature alignment; Gemini2.0-flash ranks first in model composite score, which exhibits a "language-dominant" characteristic, indicating its architecture may prioritize textual reasoning processes. Traditional architecture models generally suffer from modality bias, as seen in Qwen2-vl and Yi-vision-v2, exposing the limitations of dense parameter architectures in multimodal fusion.

According to our data, the distinction between visual and linguistic abilities is significant in many mainstream LMMs, with a performance gap of up to 16.3% between these two dimensions. However, LMMs that excel can effectively integrate both capabilities, achieving high performance in both vision and language tasks. In contrast, LMMs with weaker abilities exhibit more pronounced disparities between their visual and linguistic capabilities.

Robustness and dynamic stability indices highlights differences in anti-interference capabilities across structure design (Yang et al., 2024; Mahaut et al., 2024). ChatGlm-4v leads with a model dynamic stability score of 49.0%, potentially enhanced by its hybrid architecture to resist adversarial samples (Du et al., 2022). SSR-VLES evaluation reveals a significant negative correlation between model capability and dynamic stability of the model, with top performing models generally facing stability deficiencies. The capability leaders Doubao1.5 and Gemini2.0 achieve only 30.1%/35.5% model dynamic stability—less than half of their capability scores—while the mid-tier model ChatGlm-4v attains 49.0% model dynamic stability through its hybrid architecture, validating the potential of architectural innovation to break the "capability-model dynamic stability trade-off." Commercial model version iterations expose model dynamic stability risks, with ChatGPT4o-all showing a 2.3% decrease compared to the standard version, reflecting how parameter scaling may compromise system robustness.

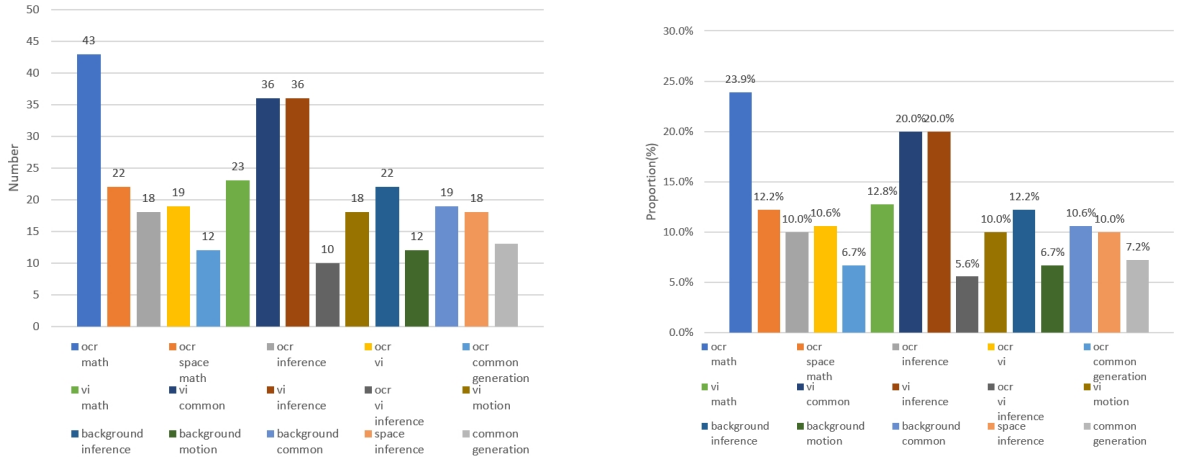

Figure. 5. This chart presents the statistical distribution of each combination. (a) shows the frequency of each label, while (b) illustrates the proportion of each label. The total percentage exceeds 100% because individual tasks may involve multiple labels.

Model	Ocr	Vi	space	Motion	Background	Common	Generation	Math	Inference	Hallucination	Input	Model composite score	Model capability
Claude3.5	65.6%	74.0%	65.8%	76.5%	76.1%	81.4%	76.2%	54.5%	54.2%	32.0%	48.5%	68.0%	70.3%
deepseek-vl2	44.9%	55.0%	42.0%	68.1%	65.2%	46.1%	13.9%	21.4%	49.2%	9.5%	12.1%	44.8%	47.9%
Doubao1.5	72.4%	79.0%	74.7%	85.4%	87.5%	85.7%	76.8%	64.7%	73.7%	14.4%	22.2%	72.0%	76.6%
Gemini2.0-flash	81.7%	78.0%	72.8%	62.9%	77.9%	82.7%	74.5%	82.6%	67.7%	12.4%	47.5%	72.3%	76.4%
ChatGlm-4v	58.6%	68.9%	56.2%	71.7%	82.4%	83.7%	85.2%	42.2%	53.9%	34.0%	53.5%	72.3%	65.6%
ChatGpt4o	67.8%	73.6%	70.7%	56.1%	76.5%	86.9%	88.9%	58.7%	71.0%	15.7%	27.3%	64.0%	73.1%
ChatGpt4o-all	61.6%	70.1%	64.1%	52.1%	67.7%	59.3%	33.7%	45.7%	49.5%	16.0%	30.3%	69.0%	58.8%
InternVL2	60.4%	71.2%	65.8%	55.7%	57.4%	58.4%	26.0%	49.3%	62.9%	32.8%	65.7%	69.0%	61.9%
Llama-3.2	59.0%	71.0%	57.6%	57.9%	79.6%	74.6%	60.2%	42.0%	58.1%	5.9%	16.7%	56.0%	62.5%
Moonshot-v1	60.1%	69.6%	61.8%	60.0%	69.2%	59.7%	22.2%	39.9%	60.6%	7.8%	10.1%	60.6%	59.0%
QVQ	74.7%	78.7%	72.2%	75.8%	69.1%	77.6%	67.2%	69.0%	57.7%	8.5%	51.5%	58.1%	69.8%
Qwen2-vl	54.1%	69.2%	60.6%	65.1%	70.8%	72.3%	40.7%	56.1%	65.3%	20.9%	27.6%	55.6%	61.8%
Yi-vision-v2	44.4%	66.2%	48.9%	75.8%	73.3%	54.8%	30.0%	36.8%	49.5%	22.9%	39.6%	66.1%	53.8%

Table 2: Independent ability score results with highest scores underlined. Model composite score includes visual, language, and robustness. The Model capability score integrates visual and language capabilities.

3.3 Independent Ability

Table 2 shows the scores of the 11 capability. These data reflect the quantitative capability of LMMs in a single function. The performance of each LMMs can be presented in a more granular manner.

Ranked the first in model capability Doubao1.5 LMMs points and individual ability to get the most times, including Vi, Space, Motion, Background, Inference, and OCR and Common ranked second. However, it scored low on both problematic robustness tests. That pushed it down to second place in overall ability, barely missing first place. This excellent capability can be found in its model architecture, which is currently more advanced MoE (Tian et al., 2024; Dai et al., 2024) architecture, with good performance in multi-task learning. Meanwhile this also exposes its poor performance in robustness and dynamic stability.

Gemini2.0-flash is the first overall ranking in model composite capability, although only two capabilities ranked first, but its many capabilities ranked at the forefront of the overall score ultimately first. The balanced development of multiple independent capabilities can make LMMs show better comprehensive performance.

3.4 Integration of Multiple Capabilities

Table 3 reflects the scores for the integration of multiple competencies. Integration of multiple capabilities refers to the simultaneous examination of multiple capabilities for a single problem. These are questions that are used in specific application scenarios and often look at various capabilities rather than a single capability. For example, when LMMs are faced with the question of the total price of all apples in the picture, they need to identify the apples in the picture, get the number of apples, and then calculate the total price of apples according to the unit price of apples given in the picture. In this process, the abilities of OCR, Vi and Math are examined respectively. Most of the problem sets we design are such comprehensive problems,

Model	math ocr	math space ocr	inference ocr	vi ocr	generation common ocr	math vi	common vi	inference vi	inference vi ocr	vi motion	background motion	background inference	common background	space inference	generation common	Combined score
Claude3.5	55.0%	56.1%	33.3%	49.1%	80.8%	63.8%	83.8%	48.1%	40.0%	73.7%	66.7%	68.2%	77.8%	49.1%	82.3%	61.6%
deepseek-vl2	26.0%	31.8%	35.2%	29.8%	10.0%	31.9%	53.7%	45.4%	33.3%	61.1%	66.7%	63.6%	65.6%	41.7%	9.2%	41.0%
Doubao1.5	65.5%	68.2%	72.2%	54.4%	85.8%	58.7%	79.6%	66.7%	70.0%	82.0%	91.7%	81.8%	99.1%	58.3%	79.2%	72.6%
Gemini2.0-flash	86.0%	81.8%	77.8%	67.5%	81.1%	73.9%	86.6%	61.1%	70.0%	58.3%	58.3%	81.8%	72.2%	47.2%	82.6%	73.9%
ChatGlm-4v	41.1%	50.0%	52.8%	41.4%	96.1%	37.7%	80.6%	47.7%	45.0%	57.2%	75.0%	75.8%	90.7%	47.2%	96.4%	59.5%
ChatGpt4o	67.4%	72.7%	64.8%	45.6%	93.1%	47.8%	91.2%	65.7%	46.7%	59.3%	47.2%	86.4%	78.7%	61.1%	93.6%	69.2%
ChatGpt4o-all	51.2%	50.0%	38.9%	12.6%	35.0%	34.8%	69.6%	41.7%	20.0%	49.4%	50.0%	59.1%	55.6%	52.8%	40.0%	46.6%
InternVL2	50.4%	54.5%	66.7%	28.4%	22.5%	42.8%	74.3%	54.5%	50.0%	48.7%	36.1%	77.3%	55.6%	60.2%	28.5%	53.0%
Llama-3.2	34.5%	26.5%	43.5%	54.7%	60.3%	43.5%	83.1%	53.2%	50.0%	38.9%	58.3%	80.3%	72.2%	32.4%	63.3%	52.8%
Moonshot-v1	43.4%	43.9%	44.4%	36.8%	21.9%	37.0%	72.0%	55.6%	40.0%	53.0%	58.3%	77.3%	66.7%	52.8%	27.9%	51.0%
QVQ	73.3%	65.9%	60.2%	49.1%	73.1%	69.6%	80.1%	50.0%	50.0%	72.2%	58.3%	63.6%	64.8%	47.2%	75.1%	64.6%
Qwen2-vl	54.4%	54.5%	61.1%	51.6%	38.9%	64.3%	83.3%	61.9%	60.0%	59.6%	66.9%	77.3%	73.3%	63.9%	43.6%	62.5%
Yi-vision-v2	35.3%	34.1%	35.6%	33.7%	34.4%	37.7%	63.6%	45.1%	34.0%	57.2%	58.3%	69.7%	69.8%	38.9%	31.8%	46.0%

Table 3: The score of the combination of various abilities of the model to be tested is counted by 100%. The highest score of a certain group of abilities in the model to be tested is indicated by underline. "Combined score" represents the average score of the various combinations.

Scoring model	Vision	Language	Model capability	Robustness	Model dynamic stability	Combined score	Model composite score
Humans	61.5%	56.7%	57.5%	19.1%	17.00%	59.1%	53.7%
DeepSeek-r1	64.2%	62.7%	61.8%	23.5%	29.7%	62.5%	58.6%
O1	51.0%	47.7%	47.7%	24.3%	22.7%	49.7%	45.4%
DeepSeek-v3	58.2%	50.1%	52.1%	11.7%	21.3%	53.%	54.2%

Table 4: The multidimensional capabilities of the model Qwen2-vl are counted at 100% using different scoring models, "Humans" represents the result of manual scoring.

so it is relatively intuitive and reasonable to judge the performance of the model in a certain scene through the integration of multiple capabilities.

Gemini2.0-flash ranked first in the integration of various capabilities, and the number of single first is the largest and obtained seven. Two of them, DouBao1.5, ranked second overall in the integration of multiple capabilities, tied for first place. Doubao1.5 has five items to obtain the first comprehensive ranking, second only to Gemini2.0-flash, and the difference is small. Another five groups are scattered among the remaining LMMs.This phenomenon may be related to differences in the different training data used by the major vendors. The difference in training data directly leads to better performance of models in specific application scenarios.

3.5 Validity Analysis Based on LLM Score

To validate the validity of the LLM-based DeepEval-R1 Scoring Framework, we scored the same result set using different methods. Through comparative analysis of scoring data on Qwen2-vl, it is found that the large model scoring system demonstrates high consistency with human evaluations in relative ranking, with visual dimension scores showing a significant positive correlation to human judgments. DeepSeek-R1 came closest to the human assessment. Furthermore, we conducte a comparison between DeepSeek-R1's performance and manually scored results obtained from other models. The analysis reveals a linear relationship between the two sets of scores, and both exhibit similar biases across all areas. This consistency will facilitate the establishment of uniform evaluation criteria.

4 Conclusions

This paper proposes an innovative multimodal evaluation framework that systematically assesses four core dimensions: visual capability, language capability, robustness, and model dynamic stability. The multi-dimensional capability index of LMMs is obtained through this evaluation framework, and the reliability of the system is verified by experiments. Benchmark tests indicate that Doubao1.5 excels in both model and visual capabilities, Gemini2.0-flash outperforms in model composite capability, ChatGpt4o leads in language proficiency, Intern VL2 shows superior robustness, and ChatGlm-4v demonstrates outstanding dynamic stability. Notably, top models demonstrate significant performance-robustness trade-offs, with robustness scores below 30% of capability metrics. Looking ahead, we will continue to refine SSR-VLES, extending its applicability to emerging LMMs and complex application scenarios.

Limitations

Data Accuracy: The benchmark tasks of SSR-VLES are manually engineered with structured

annotation frameworks, where each task instance undergoes 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, it is possible that some anomalies may still exist and will be addressed in future updates.

Data Richness: SSR-VLES'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 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 SSR-VLES are based on ChatGPT. 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 prompts for specific task types to optimize system performance.

References

Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. 2024a. 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.

Zhe Chen, Jiannan Wu, Wenhai 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. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. 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 1280–1297.

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*.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335.

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.

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.

Ji Jiaming, Qiu Tianyi, Chen Boyuan, and Yang Yaodong. 2024. (theories, techniques, and evaluation of AI alignment). In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 120–140.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. 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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. 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. 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.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: on reliability and robustness of current estimators. 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 4554–4570.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shahrzad Sayyafzadeh, Hongmei Chi, and Mark Weatherspoon. 2024. Gemini's multimodal prowess: Robust detection of adversarial patch attacks through text and image inputs. In *Proceedings of the 2024 ACM Southeast Conference, ACM SE 2024, Marietta, GA, USA, April 18-20, 2024*, pages 313–314.

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.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue summarization with mixture of experts based on large language models. 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 7143–7155.

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.

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. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, and so on. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Shangshang Yang, Xiangkun Sun, Ke Xu, Yuanchao Liu, Ye Tian, and Xingyi Zhang. 2024. Hybrid architecture-based evolutionary robust neural architecture search. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(4):2919–2934.

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*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2024. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3):64:1–64:37.

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*.