MMErroR: A Benchmark for Erroneous Reasoning in Vision-Language Models

Yang Shi[1*], **Yifeng Xie**[2*], **Minzhe Guo**[1], **Liangsi Lu**[1], **Mingxuan Huang**[3], **Jingchao Wang**[4], **Zhihong Zhu**[4], **Boyan Xu**[1†], **Zhiqi Huang**[4]
[1]Guangdong University of Technology [2]Hong Kong Baptist University
[3]Sun Yat-sen University [4]Peking University
{sudo.shiyang, evfxie, capynt, lu.liangsi.cn}@gmail.com
huangmx53@mail2.sysu.edu.cn ethanwangjc@163.com
zhihongzhu@stu.pku.edu.cn hpakyim@gmail.com zhiqihuang@pku.edu.cn

Abstract

Recent advances in Vision-Language Models (VLMs) have improved performance in multi-modal learning, raising the question of whether these models truly understand the content they process. Crucially, can VLMs detect when a reasoning process is wrong and identify its error type? To answer this, we present MMErroR, a multi-modal benchmark of 1997 samples, each embedding a single coherent reasoning error. These samples span 24 subdomains across six top-level domains, ensuring broad coverage and taxonomic richness. Unlike existing benchmarks that focus on answer correctness, MMErroR targets a process-level, error-centric evaluation that requires models to detect incorrect reasoning and classify the error type within both visual and linguistic contexts. We evaluate 12 representative VLMs, and even the best model, Gemini-3-Pro-Preview, classifies the error correctly in only 66.65% of cases, underscoring the challenge of identifying erroneous reasoning. Furthermore, the ability to accurately identify errors offers valuable insights into the capabilities of multi-modal models. Project Page: https://mmerror-benchmark.github.io

Benchmarks	Multi-Modality	Multi-Domain	Categorize
ProcessBench (Zheng et al., 2025)	✘	✘	✘
PRISM-Bench (Fang et al., 2025)	✔	✘	✘
ErrorRadar (Yan et al., 2024)	✔	✘	✘
MMErroR (ours)	✔	✔	✔

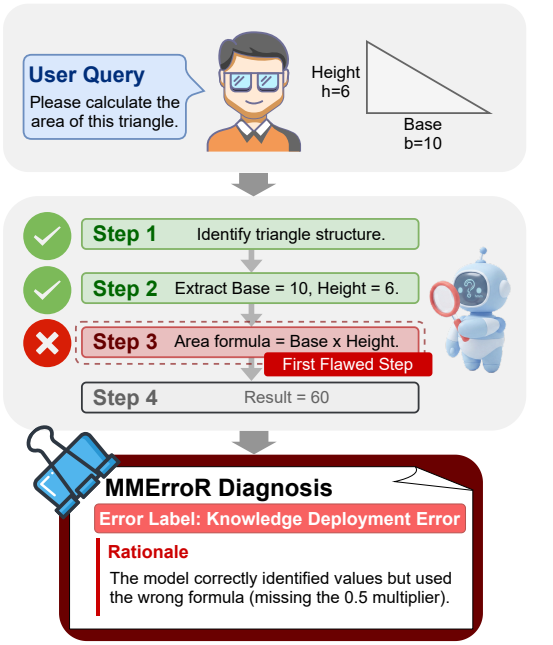

Figure 1: Comparison with existing error localization benchmarks. A sample from MMErroR illustrates an erroneous reasoning chain where the model is required to both detect and classify the error type.

1 Introduction

The rapid advancement of large multi-modal models has led to substantial progress in unified reasoning across vision and language, pushing performance (Alayrac et al., 2022; Team et al., 2023) on various multi-modal tasks closer to or surpassing in certain benchmarks (Hurst et al., 2024; Yue et al., 2024). These improvements create an impression that large multi-modal models are approaching a robust, human-like understanding of cross-modal content, a perception further reinforced by their growing deployment in real-world applications such as educational assistants, medical imaging analysis, and autonomous systems (Liu et al., 2023; Tu et al., 2024; Zitkovich et al., 2023).

Despite this progress, a fundamental question remains: Do these models genuinely understand the meaning between visual and textual content, or are they merely generating statistically plausible yet superficial associations through pattern matching? Moreover, if presented with an erroneous reasoning chain about the same multi-modal scene, can the model not only detect the error but also pinpoint its cause and type? As shown in Figure 1, existing benchmarks for error localization focus primarily

[*]These authors contributed equally to this work.
[†]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 44973–44986

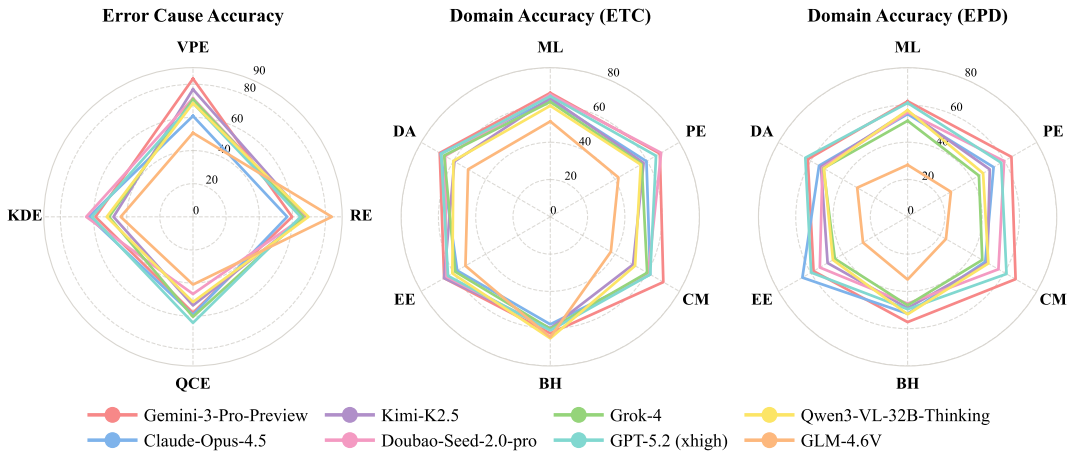

Figure 2: Comparison of different VLMs across various task domains and four error types: Visual Perception Error (VPE), Reasoning Error (RE), Question Comprehension Error (QCE), and Knowledge Deployment Error (KDE).

on identifying which step in the reasoning process is incorrect, offering limited insight into the nature of the failure. In contrast, classifying the error type enables a diagnostic understanding of why the model went astray: whether due to a breakdown in visual grounding, a logical inconsistency, a factual hallucination, or a computational mistake. Each type of error reflects a distinct weakness in the model's multimodal comprehension pipeline. Thus, a deep evaluation of a model's ability to diagnose reasoning errors serves as a litmus test for genuine multi-modal understanding.

To address this gap, we introduce MMErroR (**M**ulti-**M**odal **Error** **R**easoning Benchmark), a benchmark designed to evaluate the VLM's ability to diagnose multi-modal erroneous reasoning. MMErroR comprises 1,997 meticulously curated samples distributed across several core reasoning domains: Data & Analytics (DA), Physics & Engineering (PE), Chemistry & Materials (CM), Earth & Environment (EE), Biology & Healthcare (BH), and Mathematics & Logic (ML). Every sample contains a coherent Chain-of-Thought (Wei et al., 2022) into which a representative error has been injected. To enable unambiguous root-cause attribution, each chain is constructed to contain exactly one error. The models are required to determine whether to invoke an error label and, if so, identify its precise type. This controlled design yields fine-grained insights into model weaknesses and should be interpreted as a stress-test of error diagnosis rather than a full calibration benchmark over a natural mixture of correct and incorrect chains.

To ensure a rigorous assessment, we design two distinct evaluation modes: Error Type Classification (ETC) and Error Presence Detection (EPD). In the first mode, we explicitly inform the model that an error exists and prompt it to classify the error type. In the second mode, the model is required to first decide whether to invoke an error label before diagnosing it. Because the current release contains only erroneous reasoning chains, this second setting should be interpreted as a controlled stress-test of error sensitivity and attribution rather than a full calibration benchmark. As shown in Figure 2, the evaluation of 12 representative VLMs reveals that these tasks remain challenging. Even the most capable model in our study, Gemini-3-Pro-Preview, successfully identifies the error type in only 66.65% of cases, with performance on fine-grained error classification remaining substantially below human performance. This result underscores a notable gap between the generative capability of current models and their capacity for introspective verification.

In summary, our key contributions are as follows:

- We propose MMErroR, a benchmark designed specifically for error-type evaluation of multimodal reasoning, enabling fine-grained assessment of whether models can detect and diagnose flawed reasoning in vision-language contexts.

- Through a comprehensive empirical evaluation of 12 representative VLMs, we reveal that current models struggle significantly with introspective error detection and classification, uncovering a critical gap in their ability to

achieve trustworthy self-oversight in multi-modal reasoning.

- We conduct in-depth diagnostic analysis to uncover key factors influencing erroneous reasoning in multi-modal learning, such as modality misalignment, logical inconsistency, and perceptual over-reliance, providing actionable insights for future model improvement.

2 MMErroR

2.1 Task Classification

In MMErroR, we design two complementary evaluation tasks to assess a model's ability to detect and diagnose errors in multi-modal reasoning processes. Together, these tasks evaluate whether a model can explicitly invoke an error label under controlled erroneous conditions and, if so, correctly identify its underlying cause.

Error Type Classification (ETC) Given an image, a corresponding question, and a complete reasoning chain that is guaranteed to contain exactly one error, the model is required to identify the specific error type from a predefined taxonomy. The error types include: *Visual Perception Error*, involving incorrect visual grounding such as object misidentification, misinterpretation of spatial relations, or erroneous reading of symbols and diagrams; *Knowledge Deployment Error*, arising from misuse or misapplication of external knowledge, such as incorrect physical laws, mathematical formulas, or domain-specific concepts; *Question Comprehension Error*, caused by misunderstanding the intent of the question, overlooking key constraints, or incorrectly interpreting the required target; and *Reasoning Error*, which includes logical fallacies, missing premises, invalid inference steps, or internal inconsistencies in the reasoning process.

Error Presence Detection (EPD) Under the same input setting, the model must first determine whether to select "No Error" or "Error Present" for the provided reasoning chain. If the model predicts that an error is present, it must then determine the type of the error. Because the current release of MMErroR contains only erroneous reasoning chains, EPD is intended as a controlled stress-test of error sensitivity and attribution rather than a full calibration benchmark on mixed correct and incorrect chains.

2.2 Benchmark Construction

In this subsection, we detail the construction of MMErroR. The process is organized into four main steps.

Problem Curation To ensure both broad domain coverage and targeted evaluation of multi-modal reasoning, MMErroR sources its initial image–question–answer triplets from a set of established benchmarks, including MMMU (Yue et al., 2024), MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024), ScienceQA (Lu et al., 2022), and AI2D (Kembhavi et al., 2016). These benchmarks are widely adopted in vision–language evaluation and remain challenging for current models, providing a reliable foundation for constructing non-trivial reasoning instances.

To avoid over-representation of any single domain, we apply stratified sampling to balance the number of instances across domains. In addition, we perform a complexity-aware filtering step that removes overly simple or low-information samples, retaining only instances that require multi-step reasoning and substantive cross-modal inference. This design ensures that MMErroR emphasizes challenging reasoning scenarios rather than surface-level perception or pattern matching. Details of the filtering procedure are provided in Appendix A.

Error Injection To construct erroneous reasoning chains while maintaining control and realism, we adopt a hybrid generation strategy. For each curated instance, GPT-5 (OpenAI, 2025a) is used to inject a single, contextually coherent error into an otherwise plausible reasoning chain, under explicit generation constraints (see Appendix B). We intentionally enforce a single root-cause error per chain to obtain unambiguous diagnostic labels. Although real-world reasoning failures may involve cascaded mistakes, allowing multiple interacting errors would substantially confound attribution. The injected errors are restricted to one of four predefined categories: Visual Perception Error, Knowledge Deployment Error, Question Comprehension Error, and Reasoning Error. Aside from the injected error, the remaining reasoning steps are required to be locally coherent and logically valid, ensuring that each instance reflects a realistic and non-trivial reasoning failure.

Data Verification To ensure the quality and realism of the generated erroneous reasoning chains, we employ a rigorous three-round human verifica-

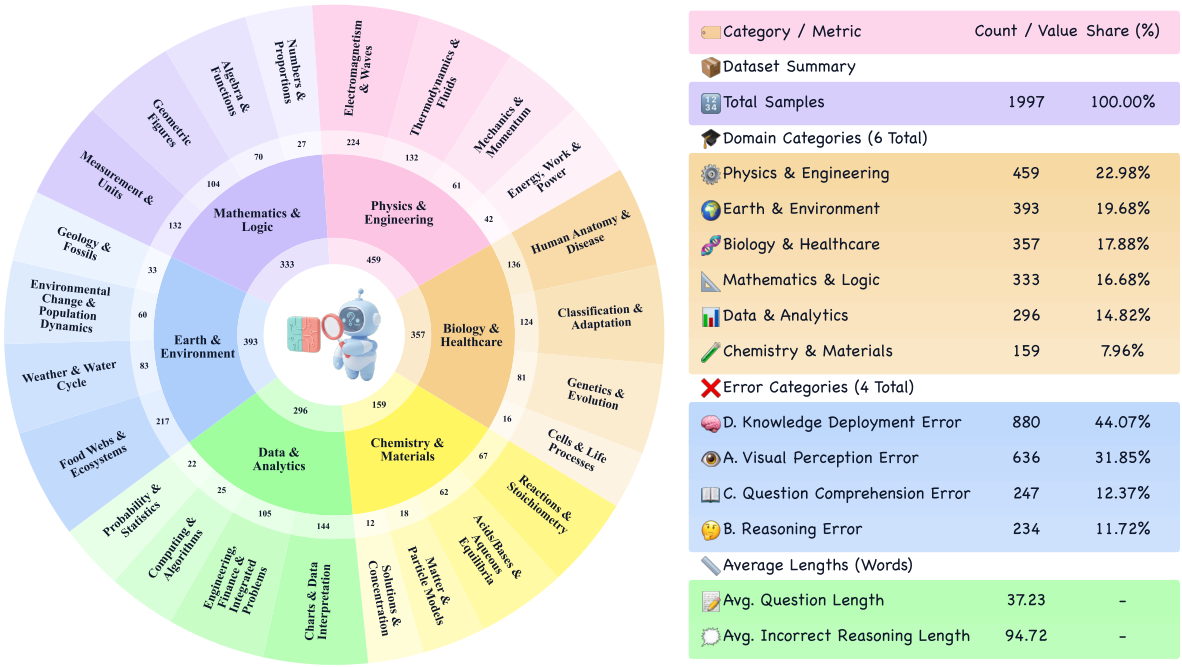

Figure 3: Detailed analysis of the domain, subdomain, and error-type statistics of MMErroR.

tion protocol. We invited a total of 20 experts (including 6 professors in the corresponding domains and 14 doctoral students) to conduct a 23-day inspection on the initial 10,000 samples. During this period, we ensured that each sample was inspected by three different experts in three separate rounds. A reasoning chain is discarded if it satisfies any of the following conditions: (1) the erroneous reasoning is incoherent or irrelevant to the original question; (2) the assigned error type is incorrect; (3) the error is ambiguous or plausibly attributable to multiple error categories. Only samples with unanimous approval are retained, resulting in 3,929 valid instances in Round 1, 3,239 in Round 2, and 3,148 instances after the third round. The marginal elimination rate of 2.81% in the final round reflects an observed agreement of 97.19% (Artstein and Poesio, 2008), suggesting annotation stability.

Quality Assurance To further ensure the quality and realism of erroneous reasoning chains in MMErroR, we apply an additional human scoring and filtering stage. Each generated reasoning chain is independently evaluated by at least two linguistics experts along four quality dimensions: *Coherence*, *Step Clarity*, *Error Localizability*, and *Semantic Consistency*. Each dimension is rated on a three-point scale: -1 (unsatisfactory), 0 (adequate), and 1 (satisfactory). A reasoning chain is retained only if its average score across evaluators exceeds a predefined threshold of 0.5. This criterion ensures that retained samples exhibit both a realistic reasoning flow and a well-localized, non-trivial error. After this scoring-based filtering, a total of 1,997 high-quality erroneous reasoning samples are retained for final inclusion. This quality assurance pipeline ensures that MMErroR is both challenging and reliable for benchmarking multi-modal error detection and diagnosis. Furthermore, a rigorous pilot study on a stratified sample of 300 instances achieved a Cohen's Kappa of $\kappa = 0.796$ (Cohen, 1960). These metrics verify the high consistency of our annotation standards.

2.3 Data Analysis

Figure 3 summarizes the hierarchical composition of MMErroR. Among the six top-level domains, Physics & Engineering accounts for the largest portion of the dataset (22.98%, 459 samples), followed by Earth & Environment (19.68%, 393 samples), Biology & Healthcare (17.88%, 357 samples), Mathematics & Logic (16.68%, 333 samples), and Data & Analytics (14.82%, 296 samples), while Chemistry & Materials constitutes 7.96% (159 samples). At the subdomain level, the benchmark remains distributed across 24 categories, with Electromagnetism & Waves, Food Webs & Ecosystems, Charts & Data Interpretation, and Human Anatomy & Disease among the largest groups. At

the error-type level, Knowledge Deployment Error is the most prevalent (44.07%, 880 samples), followed by Visual Perception Error (31.85%, 636 samples), Question Comprehension Error (12.37%, 247 samples), and Reasoning Error (11.72%, 234 samples). On average, questions contain 37.23 words, and erroneous reasoning chains average 94.72 words, indicating non-trivial reasoning contexts with multiple intermediate steps. Overall, this balanced yet challenge-oriented distribution enables MMErroR to cover diverse multi-modal scenarios while focusing on process-level reasoning failures rather than superficial answer mistakes.

3 Experiment Settings

3.1 Models

We evaluate a representative set of 12 VLMs and organize them into two groups according to model accessibility, namely open-weights models and proprietary models. The open-weights group includes LLaMA-4-Scout (Meta, 2025b), LLaMA-4-Maverick (Meta, 2025a), Qwen3-VL-32B-Instruct (Bai et al., 2025), Qwen3-VL-32B-Thinking (Bai et al., 2025), GLM-4.6V (Hong et al., 2025), and Kimi-K2.5 (Team et al., 2026). The proprietary group includes Qwen-VL-Max (Qwen Team, 2024), Grok-4 (xAI, 2025), Claude-Opus-4.5 (Anthropic, 2025), GPT-5.2 (xhigh) (OpenAI, 2025b), Doubao-Seed-2.0-pro (ByteDance Seed Team, 2026), and Gemini-3-Pro-Preview (Google, 2025). In addition to these models, we report results for simple baselines (Random Choice) and for Human Expert (Low/High) performance to assess the difficulty of MMErroR.

3.2 Implementation and Metrics

We assess model performance using two complementary evaluation protocols: Error-Type Classification (ETC) and Error Presence Detection (EPD). In both settings, each evaluation instance consists of an image, a question, and a step-by-step reasoning chain. For the ETC task, the chain is guaranteed to contain exactly one error, and the model must identify its type from four predefined categories. For the EPD task, the model must first choose between "No Error" and "Error Present" and, if it predicts an error, further classify its type. We emphasize that, because MMErroR contains only erroneous reasoning chains, EPD does not measure false-positive behavior on verified clean chains and is not intended as a full calibration benchmark. Instead, it serves as a controlled stress-test of error sensitivity and attribution under uniformly erroneous conditions. An always-error strategy does not trivially solve EPD, because credit is awarded only when the model also identifies the correct error type; such a strategy therefore reduces to ETC-level performance.

We adopt a multiple-choice format. Models are prompted to output the label corresponding to their judgment. To provide a fine-grained analysis, we report performance across six distinct dimensions: Data & Analytics (DA), Physics & Engineering (PE), Chemistry & Materials (CM), Earth & Environment (EE), Biology & Healthcare (BH), and Mathematics & Logic (ML). We also report the Macro Average Score (Macro) across these categories and the Overall Weighted Accuracy (Overall). To ensure deterministic and reproducible comparisons, we set the decoding temperature to 0 for all evaluations.

4 Empirical Results and Analysis

4.1 ETC Evaluation Results

We evaluate performance using the Error Type Classification (ETC) task. As shown in Table 1, the following observations can be made:

(1) Proprietary models achieve the strongest overall performance on ETC. Gemini-3-Pro-Preview attains the best overall accuracy at 66.65%, followed by Doubao-Seed-2.0-pro at 64.80% and GPT-5.2 (xhigh) at 64.30%. Despite this progress, the gap to Human Expert performance remains substantial, indicating that fine-grained diagnosis of erroneous reasoning is still challenging for current VLMs.

(2) Performance is heterogeneous across domains rather than dominated by a single model. Gemini-3-Pro-Preview achieves the best results in ML at 66.37% and in CM at 69.81%, Doubao-Seed-2.0-pro leads PE at 67.32%, Qwen-VL-Max performs best in BH at 66.39%, Kimi-K2.5 achieves the top score in EE at 66.67%, and GPT-5.2 (xhigh) leads DA at 69.59%. This pattern suggests that error diagnosis in MMErroR depends on multiple underlying capabilities, including domain knowledge, visual grounding, and procedural reasoning.

(3) Among open-weights models, performance also varies substantially. Kimi-K2.5 attains the strongest overall result at 60.19%, followed by Qwen3-VL-32B-Thinking at 59.29%, and both remain competitive with several proprietary models

	ML	PE	CM	BH	EE	DA	Macro	Overall
Baselines								
Random Choice	22.10	23.62	24.18	24.06	21.50	25.53	23.50	23.45
Human Expert (Low)	78.33	75.63	73.75	77.09	74.70	76.85	76.06	76.22
Human Expert (High)	91.07	88.65	87.50	90.15	88.96	90.18	89.42	89.52
Open-weights Vision-Language Models								
LLaMA-4-Scout	43.84	34.86	39.62	49.02	28.24	42.23	39.64	39.06
LLaMA-4-Maverick	42.64	36.17	39.62	50.14	28.50	42.57	39.94	39.46
Qwen3-VL-32B-Instruct	47.45	32.68	28.93	58.54	36.64	51.35	42.60	43.01
GLM-4.6V	51.05	41.18	37.11	63.87	52.16	51.35	49.45	50.23
Qwen3-VL-32B-Thinking	59.46	54.90	52.20	65.83	60.81	59.80	58.83	59.29
Kimi-K2.5	63.66	55.56	51.57	58.82	66.67	61.15	59.57	60.19
Proprietary Vision-Language Models								
Qwen-VL-Max	53.45	65.36	69.18	66.39	42.75	57.09	59.04	58.19
Grok-4	61.56	55.56	59.75	61.34	59.29	65.88	60.56	60.19
Claude-Opus-4.5	62.76	61.00	61.64	57.70	56.74	68.58	61.40	61.04
GPT-5.2 (xhigh)	64.56	63.62	62.26	60.50	65.14	69.59	64.28	64.30
Doubao-Seed-2.0-pro	65.47	67.32	61.01	59.94	66.16	66.22	64.35	64.80
Gemini-3-Pro-Preview	66.37	66.88	69.81	64.43	65.39	69.26	67.02	66.65

Table 1: Accuracy (%) comparison of baselines under ETC evaluation. The best, second-best, and third-best vision-language models in each column are highlighted by dark, medium, and light background shades, respectively. Baseline methods and human experts are excluded from ranking. Within the open-weights and proprietary groups, models are ordered by Overall accuracy in ascending order.

in selected domains. In contrast, LLaMA-4-Scout, LLaMA-4-Maverick, and Qwen3-VL-32B-Instruct remain clearly below the leading group, showing that MMErroR provides meaningful discrimination across model families and capability levels.

4.2 EPD Evaluation Results

The Error Presence Detection (EPD) task presents a more challenging setting than the ETC task, requiring models to first determine whether to invoke an error label before attempting to classify the error type. As shown in Figure 4 and Table 2, all models exhibit a clear performance drop under EPD relative to ETC, while the overall ranking remains broadly consistent. Gemini-3-Pro-Preview attains the best overall accuracy at 61.39% and the best macro-average at 61.88%, followed by GPT-5.2 (xhigh) with 58.54% overall accuracy and Claude-Opus-4.5 with 55.18%. Among open-weights models, Kimi-K2.5 performs best with an overall accuracy of 51.63%, followed by Qwen3-VL-32B-Thinking at 50.43%, showing that competitive open-weights models can retain relatively strong performance under the more challenging EPD setting. Domain-wise strengths remain diverse in EPD: Gemini-3-Pro-Preview leads ML, PE, CM, and BH, Claude-Opus-4.5 performs best in EE, and GPT-5.2 (xhigh) achieves the top score in DA.

4.3 Analysis of Reasoning Consistency

As shown in Table 3, to examine the relationship between error diagnosis and question-answering ability, we construct two evaluation subsets based on model performance in the ETC task. For each model, we randomly select 200 samples where it correctly identified the error type and 200 samples where it misidentified the error type. We then re-evaluate the same models on the original VQA task using only these two subsets. The results reveal a strong diagnosis–accuracy consistency: when the model previously diagnosed the error correctly, it also achieves significantly higher accuracy in answering the original visual question on the same subset. Conversely, samples that were misdiagnosed are strongly correlated with lower VQA accuracy. This pattern indicates that a model's ability to pinpoint what went wrong is closely tied to its underlying comprehension of the problem, which in turn supports more reliable answer generation in

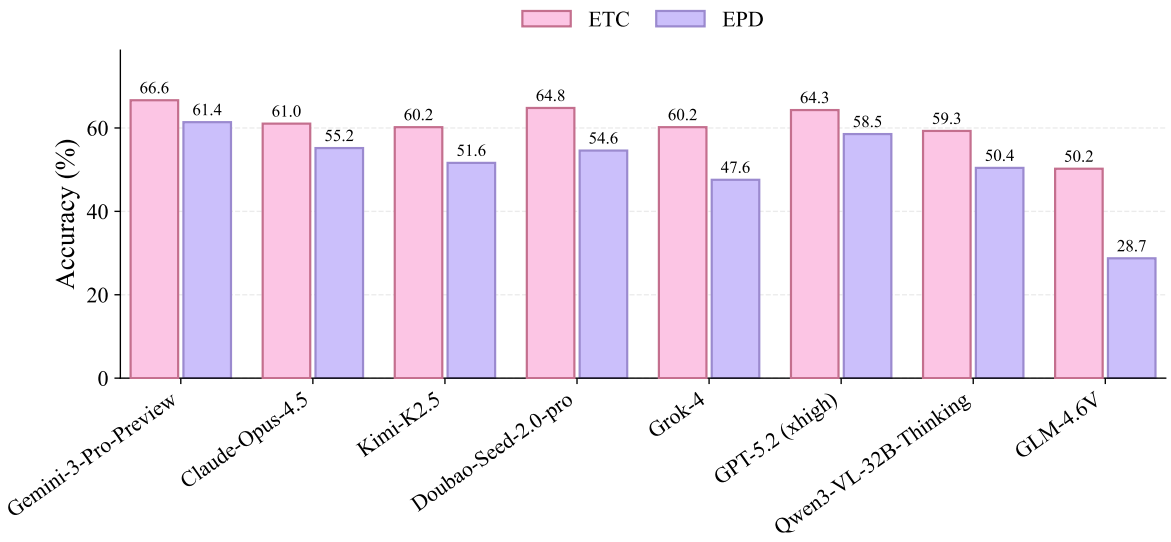

Figure 4: Performance comparison of different VLMs on MMErroR. We evaluate and compare performance under two settings: Error Type Classification (ETC) and Error Presence Detection (EPD).

the original task.

4.4 Analysis of Multi-modal Alignment

A key challenge in multi-modal reasoning is ensuring robust cross-modal alignment between visual inputs and textual descriptions (Tang et al., 2025). Inspired by (Neo et al., 2025), we selected samples from the "Visual Perception Error" category to investigate why models succeed or fail in such cases. For the Qwen3-VL-32B-Instruct model, we perform a visual analysis by extracting the logit lens of each token at each layer after the text and image inputs are processed by the VLM.

As shown in Figure 5, in case (a), where the model successfully identifies the error type, the relevant text tokens maintain a strong and correct semantic alignment with the corresponding image regions (e.g., the token "darkest cone" precisely attends to the visual cone area). In contrast, in case (b) where the model fails to detect the error, this alignment is disrupted. The model extracts irrelevant or ambiguous semantic information from the corresponding image patches (e.g., failing to associate the "arrow" token with its correct directional meaning relative to the objects).

4.5 Exploration of Steps in Reasoning

Prior research on error localization has predominantly focused on identifying which step in a reasoning chain contains an error. In this subsection, we go beyond step localization and examine how different levels of error awareness influence a model's ability to generate correct answers. We conduct this auxiliary analysis across multiple model families using a randomly sampled set of 200 examples from MMErroR. These results are intended to compare different levels of error awareness under a controlled subset setting, rather than to be directly compared with the full-benchmark results in Table 1. As shown in Table 4, we observe that merely exposing the model to the erroneous reasoning chain (*VQA+Err*) yields almost no improvement over the baseline (*VQA*). Annotating the erroneous step (*VQA+Err+StepKnown*) results in a modest but consistent performance gain across all models. The most substantial improvement, however, occurs when the error type is provided (*VQA+Err+TypeKnown*), leading to a clear and objective increase in correction accuracy. Furthermore, we observe that the effectiveness of error-type guidance varies with model capability. For the strongest models in Table 4, providing the precise error type yields the largest gains over both *VQA+Err* and *VQA+Err+StepKnown*, indicating that accurate diagnosis provides directly actionable information for answer correction.

5 Related Work

5.1 Evaluation of Multi-Modal Reasoning

The rapid evolution of Vision-Language Models (VLMs) has necessitated rigorous benchmarks to measure their progress. Initial evaluations primarily focused on simple visual question answering (VQA). More recently, comprehensive benchmarks

	ML	PE	CM	BH	EE	DA	Macro	Overall
Open-weights Vision-Language Models								
LLaMA-4-Maverick	23.12	16.12	7.55	17.09	12.21	30.41	17.75	18.13
LLaMA-4-Scout	24.62	18.74	15.09	18.21	13.23	28.72	19.77	19.73
GLM-4.6V	27.93	27.23	23.27	35.01	25.45	31.76	28.44	28.74
Qwen3-VL-32B-Instruct	40.84	36.17	27.04	44.54	27.99	41.55	36.36	36.91
Qwen3-VL-32B-Thinking	57.06	46.84	49.69	54.06	44.27	52.70	50.77	50.43
Kimi-K2.5	56.46	48.58	49.69	51.54	49.11	55.41	51.80	51.63
Proprietary Vision-Language Models								
Qwen-VL-Max	39.04	42.05	42.14	47.34	30.28	40.54	40.23	39.96
Grok-4	51.65	43.57	45.91	50.14	42.75	53.38	47.90	47.57
Doubao-Seed-2.0-pro	56.16	58.17	56.60	49.86	53.44	53.38	54.60	54.58
Claude-Opus-4.5	55.26	50.76	47.80	55.74	62.85	55.07	54.58	55.18
GPT-5.2 (xhigh)	60.96	55.99	61.64	51.26	59.29	65.88	59.17	58.54
Gemini-3-Pro-Preview	61.86	63.83	66.67	59.94	56.49	62.50	61.88	61.39

Table 2: Accuracy (%) comparison under EPD evaluation. The best, second-best, and third-best vision-language models in each column are highlighted by dark, medium, and light background shades, respectively. Within the open-weights and proprietary groups, models are ordered by Overall accuracy in ascending order.

Model	Cor.	Incor.
Gemini-3-Pro-Preview	85.5	74.5
GPT-5.2 (xhigh)	87.0	71.5
Doubao-Seed-2.0-pro	85.0	72.0
Qwen3-VL-32B-Instruct	80.5	71.0
LLaMA-4-Maverick	75.0	72.5

Table 3: Experiments on original VQA accuracy (%). For each model, "Cor." is evaluated on a randomly sampled subset of 200 examples for which the model correctly identified the error type in ETC, while "Incor." is evaluated on a randomly sampled subset of 200 examples for which the model incorrectly identified the error type.

such as MMMU (Yue et al., 2024), MathVista (Lu et al., 2023), and MathVerse (Zhang et al., 2024) have been introduced to evaluate complex reasoning capabilities across diverse domains, while related multimodal tasks have also expanded to temporally grounded video understanding (Xu et al., 2025; Ma et al., 2025a). However, these benchmarks typically adopt an outcome-oriented evaluation paradigm, focusing primarily on the correctness of the final answer. While high accuracy on these tasks suggests strong performance, it often creates an ambiguity: it is unclear whether the model genuinely understands the cross-modal content or is merely relying on superficial pattern matching. MMErroR departs from this tradition by shifting the focus from answer correctness to process-level verification. Instead of merely checking if the result is right, we evaluate whether the model can discern the validity of the reasoning path itself, providing a more transparent assessment of true multi-modal understanding.

5.2 Hallucination and Visual Consistency

Ensuring the reliability of VLMs has led to a significant body of work on hallucination detection, mitigation, and mechanism analysis. Benchmarks like POPE (Li et al., 2023) and HallusionBench (Guan et al., 2024), together with related studies on fine-grained visual perception, causal attention, shallow-layer attention repair, and attention-sink analysis (Ma et al., 2025b; Zhao et al., 2025; Zhang et al., 2025, 2026), have been instrumental in understanding and reducing object-level failures, such as errors in object existence or attribute description. While these works effectively target Visual Perception Error, they often overlook the complexity of higher-order cognitive failures. Multi-modal reasoning requires not only accurate perception but also the logical integration of visual data with parametric knowledge. As defined in our taxonomy, errors can stem from diverse sources beyond perception, including Visual Perception Error (VPE), Knowledge Deployment Error (KDE), Reasoning Error (RE), and Question Comprehension Error (QCE). MMErroR provides a broader coverage of

Model	VQA	VQA+Err	VQA+Err+StepKnow	VQA+Err+TypeKnown
Gemini-3-Pro-Preview	81.0	82.5	84.0	**90.5**
GPT-5.2 (xhigh)	80.0	80.5	82.0	**89.5**
Doubao-Seed-2.0-pro	80.5	81.5	83.0	**88.5**
Qwen3-VL-32B-Instruct	78.5	80.0	82.5	**84.5**
LLaMA-4-Maverick	73.0	74.0	75.5	**76.5**

Table 4: Impact of error awareness on correction accuracy, evaluated on a randomly sampled subset of 200 examples from MMErroR. **VQA** stands for the original VQA task, **Err** indicates that the model is additionally provided with an erroneous reasoning chain in the prompt, **StepKnown** specifies which step contains the error, and **TypeKnown** provides the exact error type classification.

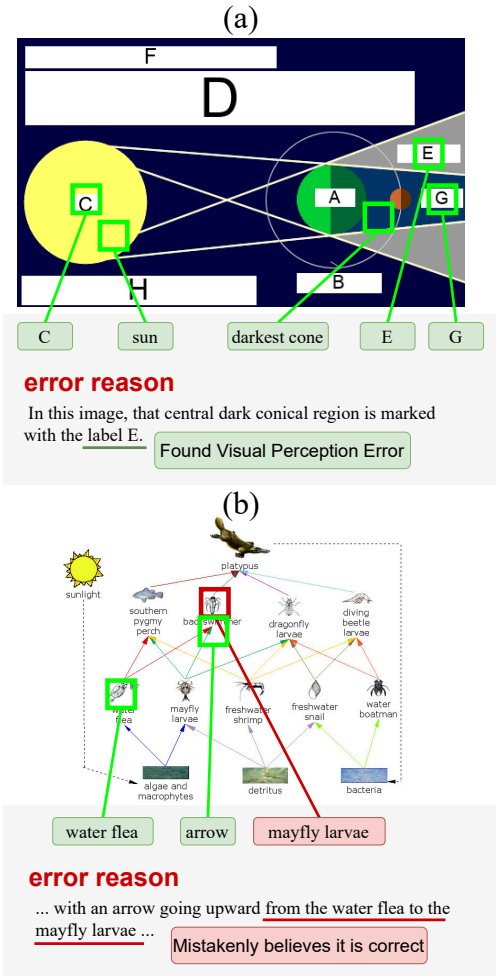

Figure 5: Logit lens visualization for image tokens. (a) Correct grounding. (b) Incorrect grounding.

these failure modes, requiring models to identify errors in logic and factual application, not just in visual grounding.

5.3 Error Localization and Erroneous Reasoning

Recent research has begun to scrutinize the intermediate steps of reasoning and the relevance of intermediate evidence to better diagnose model failures (Ruan et al., 2025; Imani et al., 2026). Existing benchmarks (Fang et al., 2025; Yan et al., 2024) represent a shift towards evaluating step-by-step consistency. These existing benchmarks primarily focus on Error Localization, identifying which step in a sequence is incorrect. While localization is useful, it offers limited insight into why the model failed. MMErroR distinguishes itself by enforcing ETC. We argue that a robust VLM must be capable of introspective diagnosis: determining whether a failure was caused by misinterpreting a diagram, applying the wrong formula, or a logical fallacy. Furthermore, beyond ETC, MMErroR also includes an EPD task, which requires models to explicitly invoke an error label before attribution under controlled erroneous conditions.

6 Conclusion

In this paper, we introduce MMErroR, a novel fine-grained benchmark designed to evaluate the reasoning capabilities of VLMs by shifting the evaluation paradigm from final-answer correctness to process-level error diagnosis. MMErroR contains 1,997 samples spanning 24 subdomains across six top-level domains and supports two core evaluation tasks: Error-Type Classification and Error Presence Detection. The current release is a controlled diagnostic stress-test built on reasoning chains with exactly one verified root-cause error. Through evaluation of 12 representative VLMs, we find that even the strongest models exhibit significant limitations in identifying and classifying reasoning errors, with the top performer achieving only 66.65% overall accuracy. We hope MMErroR can stimulate further research toward building more reliable and interpretable multi-modal reasoning systems, while verified clean chains and multi-error cascades remain important future extensions.

Limitations

Despite the comprehensive design of MMErroR, several limitations remain. First, our benchmark is constructed such that each sample contains a single, coherent reasoning error. While this isolation allows for precise diagnostic attribution, real-world reasoning failures often involve cascading or multiple simultaneous errors, which are not currently modeled in this dataset. Second, while we employ a rigorous multi-stage human verification process to ensure correctness and quality, the initial erroneous reasoning chains are generated via model-assisted synthesis. This reliance may introduce subtle biases in error patterns or linguistic styles specific to the generator model. Future work may explore open-ended generation metrics and multi-error scenarios to address these gaps.

Acknowledgments

This research was supported in part by National Science and Technology Major Project (2021ZD0111502), Natural Science Foundation of China (U24A20233, 62406078), CCF-DiDi GAIA Collaborative Research Funds (CCF-DiDi GAIA 202521), and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)(GML-KF-24-23).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anthropic. 2025. Claude opus 4.5 system card. https://www.anthropic.com/claude-opus-4-5-system-card. Accessed: 2026-04-20.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

ByteDance Seed Team. 2026. Seed2.0 model card. Official model card for the Seed2.0 series including Pro.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. 2025. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*.

Google. 2025. Gemini 3 developer guide. https://ai.google.dev/gemini-api/docs/gemini-3. Accessed: 2026-01-01.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shima Imani, Seungwhan Moon, Lambert Mathias, Lu Zhang, and Babak Damavandi. 2026. Trace: A framework for analyzing and enhancing stepwise reasoning in vision-language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3611–3625.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Hongxu Ma, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. 2025a. Ms-detr: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4514–4523.

Hongxu Ma, Chenbo Zhang, Lu Zhang, Jiaogen Zhou, Jihong Guan, and Shuigeng Zhou. 2025b. Fine-grained zero-shot object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4504–4513.

Meta. 2025a. Llama-4-maverick-17b-128e-instruct: Model card. https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct. Accessed: 2026-01-01.

Meta. 2025b. Llama-4-scout-17b-16e-instruct: Model card. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct. Accessed: 2026-01-01.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

OpenAI. 2025a. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf. Accessed: 2026-01-01.

OpenAI. 2025b. Gpt-5.2 model | openai api. https://developers.openai.com/api/docs/models/gpt-5.2. Accessed: 2026-04-20.

Qwen Team. 2024. Introducing qwen-vl. https://qwenlm.github.io/blog/qwen-vl/. Accessed: 2026-01-01.

Jiacheng Ruan, Wenzhen Yuan, Xian Gao, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Ting Liu, and Yuzhuo Fu. 2025. Vlrmbench: A comprehensive and challenging benchmark for vision-language reward models. *arXiv preprint arXiv:2503.07478*.

Zhenwei Tang, Difan Jiao, Blair Yang, and Ashton Anderson. 2025. Seam: Semantically equivalent across modalities benchmark for vision-language models. *arXiv preprint arXiv:2508.18179*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

xAI. 2025. Grok 4 (api documentation / model card). https://docs.x.ai/docs/models/grok-4. Accessed: 2026-01-01.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, and 1 others. 2024. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. 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.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3512–3534.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Jiawei Cao, Hao Cheng, and Kaijie Wu. 2026. What drives attention sinks? a study of massive activations and rotational positional encoding in large vision–language models. *Information Processing & Management*, 63(2):104431.

Qiyan Zhao, Xiaofeng Zhang, Yiheng Li, Yun Xing, Xiaosong Yuan, Feilong Tang, Sinan Fan, Xuhang Chen, Da-Han Wang, and Xu-Yao Zhang. 2025. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3981–3990.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Processbench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1024.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

A Complexity-Aware Filtering

We quantify question difficulty with a lightweight feature vector:

- comparative tokens (<, >, `taller`, `heavier`)
- negations (`not`, `never`, `except`)
- numerical quantities
- open-ended wh-words (`why`, `how many steps`)
- presence of domain-specific formulas (regex match)

Each feature is z-scored and equally weighted into a single complexity score:

$$\text{score} = \frac{1}{5}\sum_{i=1}^{5} z_i.$$

To over-sample harder instances while preserving medium-easy diversity, we fit a Gaussian $\mathcal{N}(\mu, \sigma^2)$ over all scores and draw 10 000 samples from the upper-half tail $(\mu + 0.5\sigma,\ \mu + 2\sigma)$. This raises the mean complexity from 0.00 to $+0.82$ while retaining few lower-complexity items for evaluation robustness.

B Prompt Template

To ensure transparency and reproducibility in constructing MMErroR, we detail here the prompts used to generate erroneous reasoning chains. Each prompt is carefully designed to elicit plausible yet incorrect reasoning while maintaining linguistic coherence and contextual relevance.

```
======================
ERROR TAXONOMY (CHOOSE EXACTLY ONE)
======================
1. A_Visual_Perception_Error
   The model makes a mistake in visually
interpreting the image, such as:
   - Misreading text or numbers (OCR error, e.g.,
 reading "1.0" as "10").
   - Misinterpreting chart or table values (e.g.,
 confusing bar heights).
   - Confusing colors, shapes, positions, or
object counts.
   - Mislocating objects (e.g., assigning the
wrong label to a region).
   The *reasoning logic itself* (once the wrong
visual input is assumed) should be mostly
correct.

2. B_Reasoning_Error
   The model correctly perceives the visual
information but makes a mistake in:
   - Arithmetic or calculation (e.g., 3 + 4 + 5
= 11).
   - Combining quantities, units, or proportions.

   - Logical deduction or step-by-step reasoning.

   All visually extracted facts should be
correct; the error is in the mental steps.

3. C_Question_Comprehension_Error
   The model understands the image reasonably
well but misinterprets the question, such as:
   - Answering a different but related question.
   - Ignoring constraints (e.g., "only red
objects", "in the last row").
   - Mixing up entities asked about (e.g.,
answering about Bob when asked about Alice).
   - Answering about a subset or superset
instead of the exact target.
   The reasoning may be logically consistent,
but it is applied to the wrong interpretation of
 the QUESTION.

4. D_Knowledge_Deployment_Error
   The model sees the image correctly and
understands the question, but:
   - Uses the wrong external knowledge (e.g.,
incorrect physical or scientific fact).
   - Misapplies a known formula or concept.
   - Retrieves or applies an irrelevant or
incorrect fact not supported by the image.
   Visual perception and question understanding
should be correct; the error comes from using
the wrong background knowledge or formula.

======================
TASK
======================
Given IMAGE, QUESTION, and CORRECT_ANSWER:

1. Carefully inspect the IMAGE and QUESTION.
2. Decide which single error type (A, B, C, or D
) can produce a **realistic and plausible**
incorrect answer.
3. Construct a natural, confident reasoning
chain that:
   - Uses the visual information.
   - Leads to an incorrect final answer.
   - Contains **exactly one** of the error types
 above.
   - Is otherwise as correct and detailed as
possible.
4. Do **NOT** explicitly say that you are making
 an error, simulating a failure, or referring to
 labels or taxonomy.
   - Write as if you are a normal LVLM answering
 the QUESTION.
5. Ensure the final predicted answer in `
error_reason` is **different from**
CORRECT_ANSWER.
6. Set `label` to exactly one of:
   - "A_Visual_Perception_Error"
   - "B_Reasoning_Error"
   - "C_Question_Comprehension_Error"
   - "D_Knowledge_Deployment_Error"
```

Model	0-shot	1-shot	2-shot	4-shot
Gemini-3-Pro-Preview	66.5	67.0	67.5	68.5
Doubao-Seed-2.0-pro	65.0	66.5	66.5	67.0
Qwen3-VL-32B-Instruct	49.5	53.0	55.0	56.0
LLaMA-4-Maverick	39.5	43.5	45.5	47.0

Table 5: Impact of ICL on the ETC task, evaluated on a randomly sampled subset of 200 examples from MMErroR. All 0-shot and few-shot results in this table are obtained under the same sampled subset and prompt template, and are therefore not directly comparable to the full-benchmark results in Table 1.

C Few-shot Learning Exploration

We explore whether self-oversight capabilities can be elicited or improved via In-Context Learning (ICL) (Brown et al., 2020) and few-shot prompting. To this end, we conduct an auxiliary experiment on a randomly sampled subset of 200 examples from MMErroR, and test 0-shot, 1-shot, 2-shot, and 4-shot prompts across various models, as shown in Table 5.