MMLongCite: A Benchmark for Evaluating Fidelity of Long-Context Vision-Language Models

Anonymous ACL submission

Abstract

The rapid advancement of large vision language models (LVLMs) has led to a significant expansion of their context windows. However, an extended context window does not guarantee the effective utilization of the context, posing a critical challenge for real-world applications. Current evaluations of such long-context faithfulness are predominantly focused on the text-only domain, while multimodal assessments remain limited to short contexts. To bridge this gap, we introduce *MMLongCite*, a comprehensive benchmark designed to evaluate the fidelity of LVLMs in long-context scenarios. MMLongCite comprises 8 distinct tasks spanning 6 context length intervals and incorporates diverse modalities, including text, images, and videos. Our evaluation of state-of-the-art LVLMs reveals their limited faithfulness in handling long multimodal contexts. Furthermore, we provide an in-depth analysis of how context length and the position of crucial content affect the faithfulness of these models.

1 Introduction

Recent advancements in Large Vision Language Models (LVLMs) have significantly extended their context windows, enabling them to process vast, interleaved streams of text, images and videos (Lu et al., 2024; Comanici et al., 2025; Ge et al., 2024). However, a long context window does not guarantee faithful utilization of the provided information. A critical challenge persists: when there is a conflict between the provided context and the model's internal parametric knowledge, the model often prioritizes its own knowledge, disregarding the context and leading to hallucinated responses (Tang et al., 2025). This lack of fidelity undermines the reliability of LVLMs, limiting the deployment of these models in real-world scenarios where responses need to align with context containing time-sensitive or dynamically changing information.

To enable the quantification of the faithfulness for the cutting-edge LVLMs, a prominent approach is to require models to generate citations alongside their responses (Gao et al., 2023; Hu et al., 2025). This not only compels the model to ground its output in the provided context but also empowers users with a verifiable pathway to trace the information to its source, thus facilitating a direct assessment of the response's reliability. While most citation generation benchmarks remain text-centric (Zhang et al., 2024a; Tang et al., 2025; Wu et al., 2025), the few benchmarks for multi-modal scenarios (Hu et al., 2025) suffer from significant shortcomings, including (1) **Lack Context Diversity**, as they are confined to a single document type (i.e., academic papers); (2) **Limited Task Categories**, with task formats being confined to explanation and localization, thereby failing to assess more complex cross-modal reasoning abilities; (3) **Insufficient Context Length**, failing to genuinely challenge the long-context capabilities of modern LVLMs.

To rigorously evaluate the fidelity of modern LVLMs, we introduce *MMLongCite*, a comprehensive benchmark specifically designed for long-context multimodal citation. MMLongCite is designed to overcome prior limitations by incorporating a wide diversity of data at a larger scale. To solve the first shortcoming, it incorporates three visual types-documents, images, and videos-and processes three context formats: image-only, interleaved image-text, and video-only, as shown in Figure 1. To solve the second shortcoming, it features 4 broad task categories and 8 distinct long-context tasks. Finally, to genuinely challenge modern LVLMs, context lengths are scaled from 8K to 48K tokens across 6 balanced intervals. Furthermore, to advance visual grounding and faithfulness, we introduce *MMLongCite-Grounding*. This benchmark simulates complex reasoning by presenting a composite canvas of multiple, position-indexed images. It requires models to synthesize

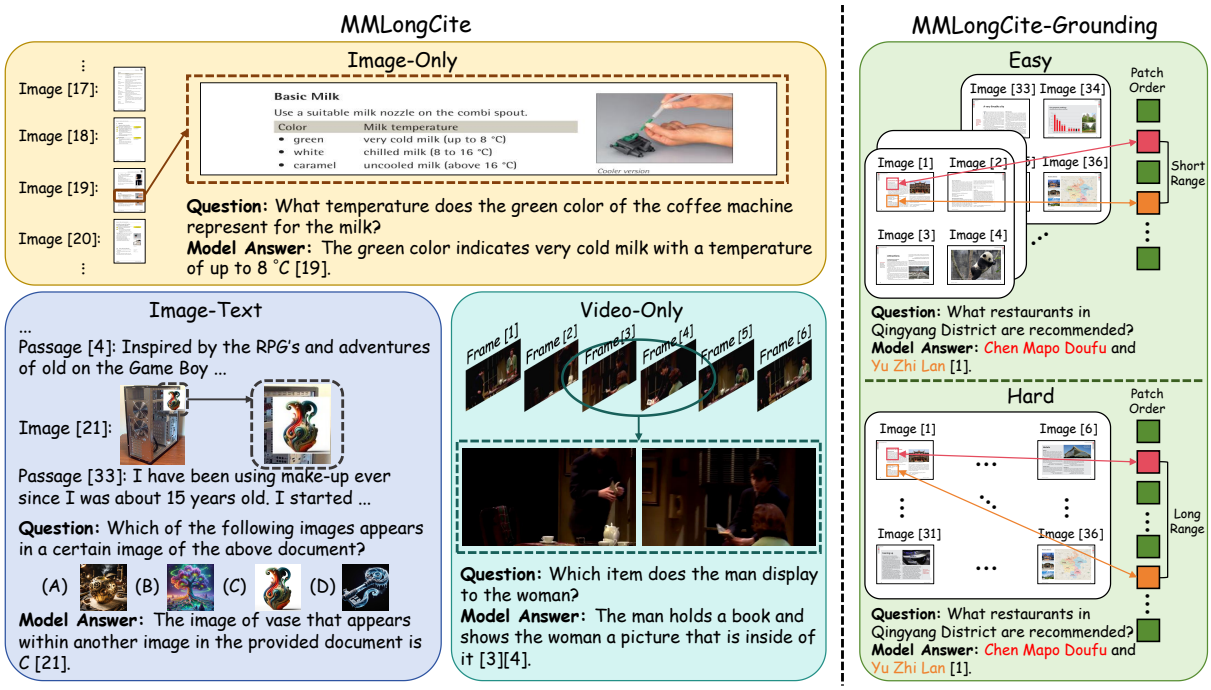

Figure 1: An overview of our proposed MMLongCite benchmark. (Left) The primary task evaluates the processing of multimodal long contexts in Image-Only, Image-Text, and Video-Only formats. (Right) MMLongCite-Grounding assesses visual grounding capabilities with scalable difficulty. In the Easy setting, models are fed a sequence of composite images, where each composite is stitched from 4 source images. In the Hard setting, models need to process a single, large composite image stitched from all source images. This often separates related evidence far apart in the patch sequence, demanding the model to resolve challenging long-range dependencies.

information across spatially distinct visual sources, mirroring real-world tasks that range from a detective piecing together clues on an evidence board, to a radiologist comparing multiple scans for a diagnosis, or an analyst monitoring a multi-panel financial dashboard.

We benchmark 12 leading LVLMs, including 10 open-source and 2 proprietary models to provide a comprehensive evaluation. Our experiments reveal a significant discrepancy between correctness and faithfulness: many models achieve high correctness scores while exhibiting poor citation performance. This finding suggests that current LVLMs often generate seemingly correct answers without faithfully grounding them in the provided context, highlighting a critical area for improvement in long-context factuality. To further investigate factors influencing faithfulness, we find that reasoning models, despite their strong correctness, surprisingly yield lower citation scores than non-reasoning models. Finally, to pinpoint the specific breaking points of model faithfulness, we conduct in-depth analyses from two perspectives: the impact of varying context lengths and the position of the key information within the long context.

2 Related Work

2.1 Faithfulness and Citation Generation

Ensuring faithfulness and mitigating hallucination remains a central challenge for large language and vision-language models. Citation generation has emerged as a prominent approach to address this issue, compelling models to ground their outputs in provided evidence and enabling verification. Benchmarks such as L-CiteEval (Tang et al., 2025), LongCite (Zhang et al., 2024a), Ref-Long (Wu et al., 2025) and ALCE (Gao et al., 2023) evaluate citation for document question answering, while frameworks like SelfCite (Chuang et al., 2025) explore retrieval-augmented citation during the generation process. However, these evaluations are confined to textual sources. The extension of citation to multimodal contexts is nascent. While MCiteBench (Hu et al., 2025) represents a pioneering step, its evaluation is constrained by a singular data domain (academic papers), a limited set of simple tasks, and context lengths that do not sufficiently tax the capacity of state-of-the-art LVLMs. Our work, MMLongCite addresses these shortcomings by introducing diverse data modalities, com-

Tasks	Source	Context	Length Distribution						Total
			0∼8k	8∼16k	16∼24k	24∼32k	32∼40k	40∼48k	
Single-Source Visual Reasoning									
LongDocURL	(Deng et al., 2025)	Image	70	70	70	70	70	70	420
MMLongBench-Doc	(Ma et al., 2024)	Image	30	30	30	30	30	30	180
Multi-Source Visual Reasoning									
HotpotQA	(Yang et al., 2018)	Text→Image	60	60	60	60	60	60	360
2WikiMultihopQA	(Ho et al., 2020)	Text→Image	60	60	60	60	60	60	360
Vision Grounding									
Visual Haystack	(Wu et al.)	Image	110	110	110	110	110	110	660
MM-NIAH	(Wang et al., 2024)	Image, Text	120	120	120	120	120	120	720
Video Understanding									
Video-MME	(Fu et al., 2025)	Video	-	20	20	20	20	20	100
LongVideoBench	(Wu et al., 2024)	Video	15	15	15	15	15	15	90

Table 1: Statistic of tasks in MMLongCite.

plex cross-modal tasks, and significantly longer contexts to establish a more comprehensive and rigorous standard for multimodal faithfulness.

2.2 Long-Context Multimodal Benchmark

The rapid expansion of model context windows has spurred the development of benchmarks aimed at evaluating long-sequence processing. In the text domain, a rich set of benchmarks (Shaham et al., 2023; Li et al., 2024; Bai et al., 2024a,b; An et al., 2024; Hsieh et al.; Dong et al., 2024; Zhang et al., 2024b; Ye et al., 2025) has been established to assess capabilities across tasks like long-document question answering, summarization, and code analysis. More recently, this focus has extended to the multimodal domain, with emerging benchmarks (Ma et al., 2024; Wang et al., 2024; Wu et al.; Wang et al., 2025a; Deng et al., 2025; Wu et al., 2024; Fang et al., 2024; Wang et al., 2025b) targeting complex abilities such as retrieving information from long interleaved sequences of text and images or understanding lengthy video narratives. However, a critical limitation pervades these existing evaluation suites: they typically assess performance based solely on the correctness of the final generated output. This approach fails to verify whether a model genuinely grounds its response in the provided context or merely recites parametric knowledge, a vulnerability that is particularly concerning given the potential for data contamination. Consequently, such evaluations may not accurately reflect the true long-context capabilities of a model (Yen et al., 2024; Tang et al., 2025). To address this gap, we introduce MMLongCite, the first benchmark designed to enforce faithfulness by integrating a core citation generation task within extensive multimodal contexts.

3 MMLongCite

Our benchmark, MMLongCite, comprises four distinct tasks: Single-Source Visual Reasoning, Multi-Source Visual Reasoning, Vision Grounding and Video Understanding. It integrates 8 diverse datasets and encompasses 2,890 samples, with the length of the context ranging from 8K to 48K. The statistics of dataset can be viewed in Table 1.

3.1 Problem Definition

Given the multimodal long context D and a query Q, the model is required to generate the response R, which is composed of n statements $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$ and their corresponding citations $\mathcal{C}_i = \{c_{i,1}, c_{i,2}, \cdots\}$. To enable the model to generate citations, we assign a unique citation index to each image, frame, or passage within the context.

3.2 Task Setup and Benchmark Construction

We construct MMLongCite by curating and adapting instances from 8 publicly available datasets. The construction process is organized into three categories based on the context modality, mirroring the structure presented in Figure 1: image-only, interleaved image-text, and video-only contexts. A detailed description is provided in Table 1.

3.2.1 Image-Only Contexts

This category evaluates a model's ability to process long contexts composed entirely of images. It comprises five datasets: LongDocURL (Deng et al., 2025), MMLongBench-Doc (Ma et al., 2024), HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), and Visual Haystack (Wu et al.).

The construction methodology varies depending on whether the task involves a single, continuous document or a collection of discrete images.

For tasks based on a single, cohesive document (LongDocURL and MMLongBench-Doc), the goal is to assess deep comprehension within a continuous visual narrative. To ensure a balanced distribution across various context lengths, we normalize each source document to a target length L_t. To preserve the original positional bias of the key information, which is known to influence model performance (Liu et al., 2024), our resizing strategy proportionally trims the prefix and suffix context surrounding the evidence region. The new lengths for the prefix (L'_p) and suffix (L'_s) are calculated as:

$$\begin{aligned} L'_p &= (L_t - L_e) \times \frac{L_p}{L_p + L_s} \\ L'_s &= (L_t - L_e) \times \frac{L_s}{L_p + L_s}, \end{aligned} \tag{1}$$

where L_e, L_p, and L_s represent the lengths of the evidence, original prefix, and original suffix, respectively.

For tasks requiring information aggregation from discrete sources (HotpotQA, 2WikiMultihopQA, and Visual Haystack), we create long contexts by placing evidence images within a long sequence of distractor images. For the text-native datasets HotpotQA and 2WikiMultihopQA, we first convert all text documents (both evidence and distractors) into images. This is achieved through a two-step pipeline: (1) we programmatically generate standardized PDF documents from raw text using the `ReportLab`[1] library, and (2) we render these PDFs into images using `PyMuPDF`[2], cropping any excess margins.

3.2.2 Interleaved Image-Text Contexts

This category, represented by MM-NIAH (Wang et al., 2024), tests a model's ability to process long, heterogeneous contexts where images and text passages are interspersed. The core task is to retrieve a specific piece of information (an *needle*), which can be either an image or a text snippet, from this mixed-modality sequence. We adapt instances from the `retrieval-image` and `retrieval-text` subtasks. The construction process involves embedding the target needle within a long context of distractor images and text blocks, ensuring a uniform distribution of samples across different context lengths.

[1]https://pypi.org/project/reportlab/

[2]https://pymupdf.readthedocs.io/en/latest/

Model	Active Param.	Total Param.	Reasoning
Proprietary Models			
Gemini-2.5-Pro (Comanici et al., 2025)	🔒	🔒	✓
Doubao-Seed-1.6 (ByteDance Seed, 2025)	🔒	🔒	✓
Open-Source Models			
Qwen2.5-VL-3B (Bai et al., 2025)	3B	3B	✗
Qwen2.5-VL-7B (Bai et al., 2025)	7B	7B	✗
MiMo-VL-7B-SFT (Team et al., 2025a)	7B	7B	✓
MiMo-VL-7B-RL (Team et al., 2025a)	7B	7B	✓
Gemma3-12B (Team et al., 2025b)	12B	12B	✗
Kimi-VL-A3B-Instruct (Team et al., 2025c)	3B	16B	✗
Kimi-VL-A3B-Think (Team et al., 2025c)	3B	16B	✓
Qwen2.5-VL-32B (Bai et al., 2025)	32B	32B	✗
Qwen2.5-VL-72B (Bai et al., 2025)	72B	72B	✗
GLM-4.5V (Hong et al., 2025)	12B	108B	✓

Table 2: Statistic of LVLMs, where 🔒 denotes proprietary model.

3.2.3 Video-Only Contexts

This category focuses on temporal reasoning and is built upon Video-MME (Fu et al., 2025) and LongVideoBench (Wu et al., 2024). A primary challenge with long videos is that the high number of frames (e.g., at 30 frames per second) generates sequences that vastly exceed the context capacity of current models. To address this, we standardize all videos by downsampling them to 1 frame per second (fps) using `FFmpeg`[3]. From these downsampled frame sequences, we then sample instances of varying lengths to construct the final evaluation set.

3.2.4 MMLongCite-Grounding

To specifically assess the visual grounding and spatial reasoning of models, we introduce MMLongCite-Grounding. This extension challenges models to operate on composite images and is structured into two difficulty levels, as shown in Figure 1 (Right). In the **Easy** setting, we merge every four context images into a single, larger composite image. In the **Hard** setting, we stitch all context images into a single, expansive canvas. Unlike the main benchmark where images are discrete inputs, MMLongCite-Grounding requires the model to first segment the composite image, identify the content within each sub-region, and link it back to its original index for citation. This design enables a more rigorous evaluation of a model's spatial reasoning, information attribution, and its ability to process visually dense and complex layouts.

Models	Single-Source Visual Reasoning				Multi-Source Visual Reasoning				Vision Grounding				Video Understanding			
	CP	CR	F_1	Cor	CP	CR	F_1	Cor	CP	CR	F_1	Cor	CP	CR	F_1	Cor
Proprietary LVLCMs																
Gemini-2.5-Pro	**92.64**	**91.60**	**91.57**	**76.63**	**96.23**	**97.00**	**96.23**	**87.01**	**88.81**	**90.59**	**89.07**	**85.72**	**83.02**	**77.00**	**77.84**	**73.94**
Doubao-Seed-1.6(w/o think)	80.04	78.66	78.39	70.38	78.02	83.34	78.10	80.62	80.73	77.57	78.50	81.09	71.81	68.18	67.01	67.47
Open-source LVLCMs																
Qwen2.5-VL-3B	21.17	24.99	21.42	46.89	11.19	12.77	11.17	42.01	12.79	10.06	10.68	47.79	37.46	40.18	36.47	61.58
Qwen2.5-VL-7B	55.33	57.35	53.63	59.12	35.59	38.16	35.43	53.61	22.24	18.11	19.03	56.72	52.06	44.78	45.05	65.94
MiMo-VL-7B-SFT	68.91	50.55	54.97	66.85	70.31	62.82	64.11	64.65	49.68	39.51	41.95	64.51	64.32	29.89	38.11	70.00
MiMo-VL-7B-RL	80.03	50.67	58.15	71.57	73.08	62.26	64.32	67.92	**68.36**	45.32	51.30	69.37	67.81	25.23	34.43	71.64
Gemma3-12B	62.41	63.83	62.58	49.24	63.34	66.15	64.27	60.07	41.24	39.30	39.64	59.03	68.04	**76.39**	**71.17**	63.50
Kimi-VL-A3B-Instruct	62.23	57.43	57.02	61.63	50.48	43.55	45.50	55.21	21.84	19.78	20.37	62.55	7.60	5.88	6.27	61.31
Kimi-VL-A3B-Think	53.98	29.82	36.11	56.85	38.00	19.26	23.62	49.38	31.66	17.68	21.33	42.16	42.86	16.48	22.17	55.36
Qwen2.5-VL-32B	70.91	51.88	57.51	66.06	53.13	44.19	46.48	66.11	55.49	36.85	42.54	68.93	64.02	34.42	41.83	73.44
Qwen2.5-VL-72B	**81.37**	**79.01**	**78.97**	67.96	56.57	58.13	56.36	65.42	65.00	**58.54**	**60.58**	73.29	**73.71**	60.14	63.50	68.94
GLM-4.5V(w/o think)	78.54	68.97	71.16	68.33	68.47	70.47	66.99	74.10	56.18	46.56	49.21	**76.79**	70.89	46.07	52.73	73.69
GLM-4.5V(w/ think)	80.45	67.23	70.17	**74.87**	**76.50**	**73.90**	**71.50**	**83.19**	60.64	40.17	45.73	74.95	63.79	33.07	41.17	**76.36**

Table 3: Performance of LVLMs in MMLongCite.

Models	LongDocURL				HotpotQA				Visual Haystack				Video-MME			
	Easy		Hard		Easy		Hard		Easy		Hard		Easy		Hard	
	F_1	Cor	F_1	Cor	F_1	Cor	F_1	Cor	F_1	Cor	F_1	Cor	F_1	Cor	F_1	Cor
Gemini-2.5-Pro	**56.73**	59.29	**34.58**	51.67	**91.99**	**93.06**	**28.18**	**80.97**	**83.71**	**72.12**	**40.68**	**65.98**	**60.74**	**78.00**	**57.39**	72.00
Qwen2.5-VL-3B	11.04	42.14	11.02	34.40	6.50	47.22	5.73	39.31	10.84	32.88	6.33	42.73	19.40	64.00	10.12	60.00
Qwen2.5-VL-7B	24.93	56.43	15.42	47.02	21.09	69.58	8.07	49.44	18.68	53.48	11.84	52.27	35.32	67.50	30.81	67.50
Qwen2.5-VL-32B	23.72	65.36	11.15	49.29	25.19	81.53	10.58	52.08	25.87	55.45	18.56	58.26	31.15	69.00	24.55	66.50
Qwen2.5-VL-72B	39.75	**65.83**	19.69	**53.21**	33.47	79.44	11.33	68.33	53.31	50.45	30.23	51.59	50.61	**78.00**	41.75	**73.00**

Table 4: Performance of LVLMs in MMLongCite-Grounding.

4 Experiment

4.1 Experimental Setup

We conducted a comprehensive evaluation of 12 recent LVLMs, comprising 10 open-source and 2 proprietary models, as detailed in Table 2. The evaluated models exhibit considerable diversity in their specifications: context windows(64K to 1M tokens), model parameter (3B to 108B), architectures (including both Dense and MoE designs) and distinct reasoning paradigms (such as direct-generation, deliberate, and automatic modes).

Following Gao et al. (2023); Tang et al. (2025); Zhang et al. (2024a), we evaluate our model on two key aspects: citation quality and generation quality. To measure citation quality, we report citation precision (CP), recall (CR), and the resulting citation F1 (F1). To evaluate generation quality, we measure correctness (Cor) by scoring each response on a 0/0.5/1 scale (incorrect/partially correct/fully correct) against the ground truth. For both evaluation aspects, we leverage the powerful GPT-4.1[4] as the adjudicator to ensure higher accuracy, compared to prior work (Gao et al., 2023; Tang et al., 2025). A detailed description of the evaluation metrics is available in Appendix B.

[3]https://ffmpeg.org/

[4]https://openai.com/index/gpt-4-1/

4.2 Main Results

We present the all evaluation results on MMLongCite in Table 3 and the model performance on MMLongCite-Grounding in Table 4.

4.2.1 Analysis of MMLongCite

Performance of Open-source Models Our investigation yields three key insights. 1) We observe that well-trained, medium-sized models like the MiMo-VL series can match or even surpass significantly larger 72B models on specific tasks. This finding underscores that a sophisticated training strategy and high-quality data can be more critical than sheer parameter count (Team et al., 2025a). 2) Experiments with the Qwen series show a clear trend where citation quality improves with increasing model parameters; thus, model scaling remains an effective strategy for enhancing this the citation capability of models. 3) Our results reveal the evidence of model specialization. Gemma3-12B, for example, achieves a standout citation F1 score in Video Understanding despite its otherwise average performance. This suggests that certain models are architecturally or functionally predisposed to excel in specific domains, irrespective of their broad-spectrum performance.

Open-source Models vs. Proprietary Models There is a significant performance gap between

open-source and proprietary models, particularly in in-context retrieval tasks. This disparity is most pronounced on the task of Vision-Grounding, which requires LVLMs to precisely retrieve information from long contexts. On this task, the top-performing open-source model (Qwen2.5-VL-72B) lags behind its leading proprietary counterpart (Gemini-2.5-Pro) by a substantial 28 points in citation F1. Furthermore, we observe that proprietary models tend to exhibit a balanced profile between citation precision and recall. In contrast, some open-source models often display a significant imbalance, frequently achieving high precision at the cost of substantially lower recall.

Impact of Reasoning Mode The results between models with reasoning and without reasoning mode highlight that Chain-of-Thought (CoT) acts as a double-edged sword in generating answers with citations. As evidenced by the performance of GLM-4.5V enabling the reasoning mode significantly enhances answer correctness and citation precision across most tasks. However, this gain comes at the cost of a general decline in citation recall. This suggests a fundamental trade-off: the CoT mechanism promotes more deliberative reasoning, boosting the quality of the final answer and the precision of its cited evidence, but it also induces a "conservative" citation behavior, causing the model to overlook some relevant evidence. Furthermore, we observe that MiMo-VL-7B-RL consistently demonstrates superior citation quality compared to its SFT counterpart. This may be partially attributed to the reinforcement learning stage incorporating high-quality, long-context grounding data. (Team et al., 2025a).

Correlation Between Citation Quality and Correctness We identify a decoupling of citation quality and answer correctness in small models (i.e., models with fewer than 7B parameters). For instance, Qwen2.5-VL-3B achieves 42.01 in correctness on Multi-Source Visual Reasoning but only 11.17 in citation F1. This indicates that small models may rely on its parametric knowledge rather than the provided context, and fail to supply supporting evidence, which is consistent with the observations by Tang et al. (2025) and Intel (2024). The phenomenon underscores that correctness alone is an insufficient metric for evaluating LVLMs.

4.2.2 Analysis of MMLongCite-Grounding

To comprehensively evaluate visual grounding capabilities under challenging conditions, we analyze the performance on MMLongCite-Grounding in Table 4. The most striking and consistent observation is a significant decline across all models when transitioning from "Easy" to "Hard" setting, as reflected in both citation F1 and correctness metrics.

Grounding Breakdown in Dense Visual Contexts The primary difficulty lies in degraded visual grounding performance. As shown in Table 4, citation F1 scores of all models drop substantially in the "Hard" setting. For example, Gemini-2.5-Pro achieves an F1 of 91.99 in HotpotQA-Easy, which sharply reduces to 28.18 in HotpotQA-Hard, a loss of over 60 points. Similarly, the citation F1 for Qwen2.5-VL-72B decreases from 39.75 to 19.69 in LongDocURL when moving to the harder setup. These results indicate significant challenges in fine-grained localization within dense visual inputs, where models struggle to associate answers with precise regions.

Comprehension Disruption from Visual Noise The drop in grounding accuracy is mirrored by a decline in answer correctness, revealing that dense visual information introduces not only grounding challenges but also distractions that impair the comprehension of models. For instance, answer correctness for Qwen2.5-VL-72B in HotpotQA-Hard declines to 68.33, compared to 79.44 in the easy setting. This suggests that visual noise hinders the ability to focus on task-relevant content, which in turn compromises question answering accuracy.

Contrasting Strengths in Grounding Versus Knowledge When comparing proprietary and open-source models, notable differences emerge. Gemini-2.5-Pro consistently outperforms open-source alternatives in visual grounding, particularly in citation F1 scores. For example, in HotpotQA-Hard, Gemini-2.5-Pro achieves an F1 of 28.18, far exceeding the F1 of 11.33 observed for Qwen2.5-VL-72B, pointing to stronger spatial reasoning and localization in complex visual layouts.

However, an interesting distinction arises in answer correctness. Despite weaker grounding performance, Qwen2.5-VL-72B occasionally matches or surpasses Gemini-2.5-Pro in answer accuracy. For instance, in LongDocURL-Hard, Qwen2.5-VL-72B achieves an answer correctness of 53.21 compared to 51.67 for Gemini-2.5-Pro, suggesting dif-

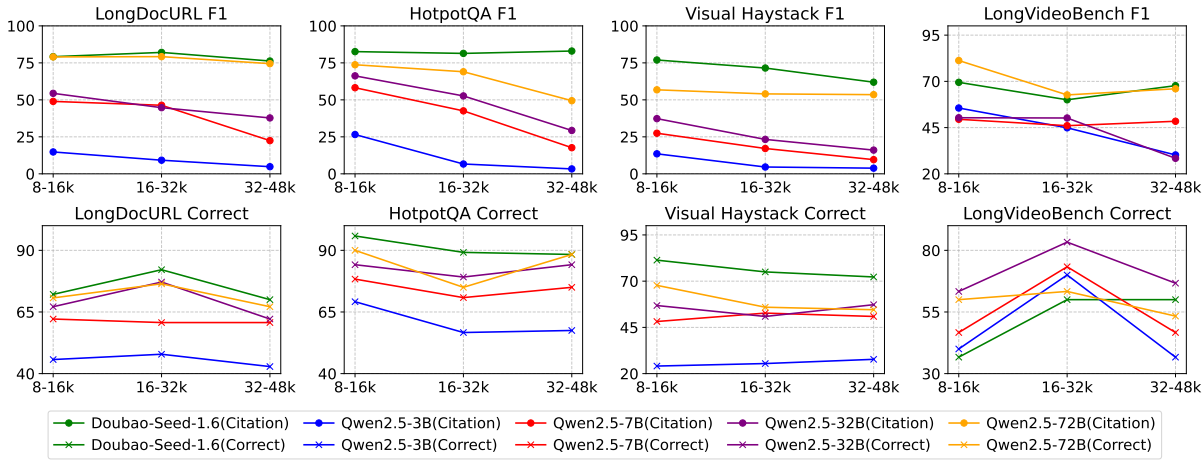

Figure 2: Model Performance across different context length intervals.

ferent underlying strategies. Gemini-2.5-Pro relies more on direct grounding and visual evidence, while Qwen2.5-VL-72B appears to leverage parametric knowledge for inference, allowing accurate answers even when grounding is incomplete.

5 Ablation Study

In this section, we conduct a series of analyses to further probe the model's capabilities in handling multimodal long contexts. Specifically, we investigate three key aspects: (1) the impact of varying context lengths on model faithfulness in § 5.1; (2) the effect of retrieval-based context compression on overall performance in § 5.2; (3) the performance variation of LVLMs with respect to the position of key information within the context in § 5.3.

Due to its distinctive strengths, including broad scalability, advanced long-context capabilities, and competitive performance against proprietary models, we select the Qwen2.5-VL series as the representative open-source model for our ablation study.

5.1 Analysis of different length intervals

To evaluate the impact of increasing context length, we analyze the model performance across 16k-token intervals, focusing on answer correctness and citation F1 in Figure 2. The results reveal that open-source Qwen2.5-VL models experience significant performance degradation as the context length grows, with smaller variants like Qwen2.5-VL-3B and Qwen2.5-VL-7B struggling to retrieve relevant information from extended inputs. Larger models, such as Qwen2.5-VL-72B, are relatively more robust but still show a notable decline in citation F1. In contrast, Doubao-Seed-1.6 exhibits exceptional stability, consistently maintaining high performance across all context lengths. These findings underscore two key insights: (1) a widening gap between generating correct answers and grounding them with precise citations; and (2) effective utilization of long contexts remains a critical bottleneck, as extending the context window can not ensure reliable reasoning over larger spans.

Model	LongDocURL		HotpotQA		Visual Haystack		Video-MME	
	F_1	Cor	F_1	Cor	F_1	Cor	F_1	Cor
Gemini-2.5-Pro	89.93	82.74	96.63	94.58	90.04	87.27	74.16	79.00
+ Retrieval	85.65	52.02	95.36	87.50	85.71	52.05	76.07	63.50
Qwen2.5-VL-3B	13.31	45.12	15.18	59.17	10.37	25.30	32.31	66.50
+ Retrieval	17.61	35.83	45.87	75.56	37.79	22.58	24.98	50.50
Qwen2.5-VL-7B	45.50	60.24	42.09	72.22	21.30	51.21	45.37	68.00
+ Retrieval	46.06	44.29	61.58	80.69	31.02	49.47	41.85	54.50
Qwen2.5-VL-32B	50.59	68.45	51.73	81.67	27.65	55.23	35.76	73.00
+ Retrieval	46.82	47.14	71.59	84.17	35.70	43.48	35.47	60.00
Qwen2.5-VL-72B	77.12	70.60	64.95	83.47	56.31	60.61	61.43	74.00
+ Retrieval	73.28	46.79	77.19	86.39	70.72	41.89	59.30	56.50

Table 5: Model Performance with the RAG method.

5.2 Analysis of Retrieval Augmented Generation

We employ clip-vit-base-patch32[5] as the embedding model within Retrieval-Augmented Generation (RAG) to retrieve images most relevant to the query. Specifically, the top 8 images with the highest relevance scores are selected as the context.

As shown in Table 5, our evaluation of RAG reveals a clear performance divergence based on task complexity. For OCR-intensive tasks like HotpotQA, which primarily rely on extracting textual information, RAG consistently enhances both correctness and citation F1 across all models. This demonstrates its efficacy in grounding

[5] https://huggingface.co/openai/clip-vit-base-patch32

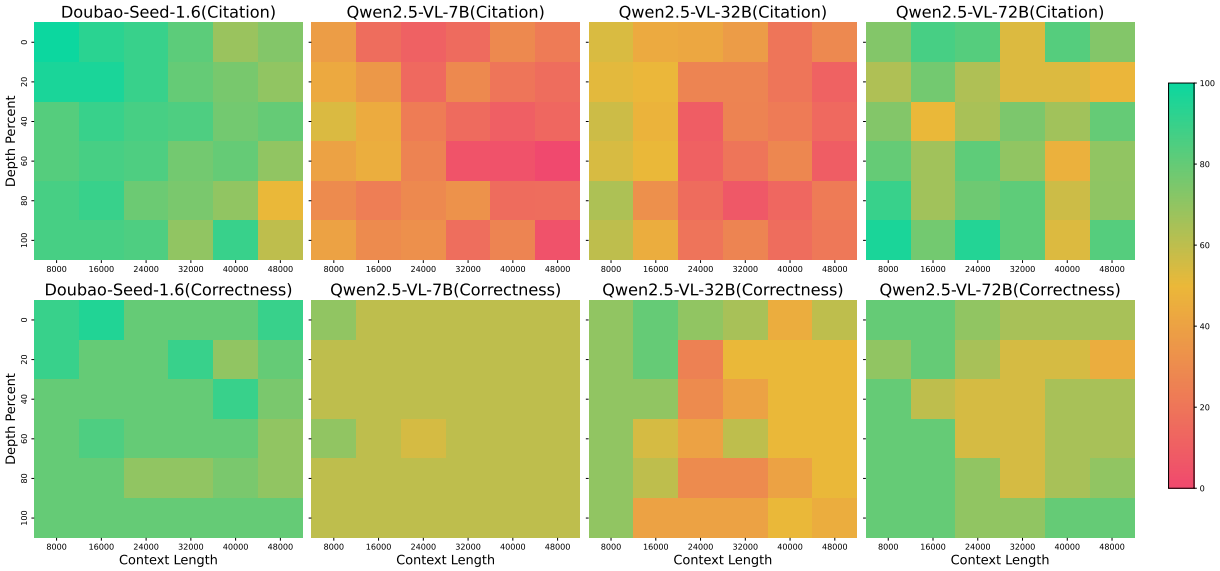

Figure 3: Model Performance on citation quality and correctness across different depths and context lengths in Visual Haystack.

responses when dealing with images of text-based content. Conversely, on benchmarks requiring nuanced visual understanding (LongDocURL, Visual Haystack, and Video-MME), the impact of RAG is mixed. While it can guide smaller models by compressing the context, it frequently degrades the performance of more powerful, larger models. This suggests that for highly capable models, the noise and potential inaccuracies from imperfect multimodal retrieval can outweigh the benefits of context reduction. The retriever may fail to capture critical visual details or introduce irrelevant data, thereby misleading the generation process and undermining the model's inherent long-context comprehension abilities.

5.3 Analysis of the Position of Needles

we present the model performance on Visual Haystack across different context lengths and needle depths in Figure 3. The heatmaps reveal two critical challenges for current LVLMs.

First, we identify a clear "lost-in-the-middle" problem. For most Qwen2.5-VL variants, performance across both correctness and citation declines sharply when the target image resides in the central 40-60% of the context depth. Second, we observe a notable disparity between the ability to answer correctly and the ability to cite accurately. The superior performance on correctness (bottom row) versus citation (top row) suggests that while models can often access the necessary information to formulate an answer, they struggle with the fine-grained positional reasoning required for precise source attribution. We also observe a clear trend where performance improves with model size, suggesting that scaling is a key factor in mitigating these challenges. For instance, in contrast to the significant performance degradation seen in the 7B model, Qwen2.5-VL-72B maintains high performance on both metrics, approaching the robustness demonstrated by Doubao-Seed-1.6.

6 Conclusion

In this work, we introduce MMLongCite, a comprehensive benchmark designed to rigorously evaluate the fidelity of LVLMs in multimodal long-context scenarios. Our extensive evaluation of state-of-the-art models reveals a significant gap between their extended context capacity and their ability to faithfully ground responses in provided evidence. A key finding is the consistent decoupling of answer correctness from citation fidelity, where models often generate correct answers without accurate source attribution, suggesting a reliance on parametric knowledge over contextual evidence. Furthermore, our in-depth analyses pinpoint specific failure modes, including performance degradation with increasing context length and visual density, a "lost in the middle" vulnerability. Collectively, these findings underscore that merely extending the context window is insufficient and highlight the urgent need for future research to focus on improving the core mechanisms for reliable information utilization and attribution in LVLMs.

Limitation

In this paper, we introduce MMLongCite, a benchmark designed to rigorously assess the faithfulness of Large Vision-Language Models within long-context multimodal scenarios. The current scope of MMLongCite, however, primarily centers on visual-centric modalities, namely images, videos, and interleaved text. Looking ahead, we plan to extend our benchmark to encompass a broader spectrum of data types, with the ultimate goal of establishing a truly pan-modal citation benchmark in a future version.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024a. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

ByteDance Seed. 2025. Seed 1.6: Tech introduction. Accessed on September 28, 2025.

Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen-tau Yih. 2025. Selfcite: Self-supervised alignment for context attribution in large language models. *arXiv preprint arXiv:2502.09604*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. 2025. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1135–1159, Vienna, Austria. Association for Computational Linguistics.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2086–2099.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Junqi Ge, Ziyi Chen, Jintao Lin, Jinguo Zhu, Xihui Liu, Jifeng Dai, and Xizhou Zhu. 2024. V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. *arXiv preprint arXiv:2412.09616*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

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-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Caiyu Hu, Yikai Zhang, Tinghui Zhu, Yiwei Ye, and Yanghua Xiao. 2025. Mcitebench: A benchmark for multimodal citation text generation in mllms. *arXiv e-prints*, pages arXiv–2503.

Intel. 2024. Do smaller models hallucinate more? *blog*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 11:157–173.

Yujie Lu, Xiujun Li, Tsu-Jui Fu, Miguel Eckstein, and William Yang Wang. 2024. From text to pixel: Advancing long-context understanding in mllms. *arXiv preprint arXiv:2405.14213*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989.

Zecheng Tang, Keyan Zhou, Juntao Li, Baibei Ji, Jianye Hou, and Min Zhang. 2025. *L-CiteEval*: A suite for evaluating fidelity of long-context models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5254–5277, Vienna, Austria. Association for Computational Linguistics.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025a. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025b. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025c. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. 2025a. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3221–3241.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, and 1 others. 2024. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, and 1 others. 2025b. Mmlongbench: Benchmarking long-context vision-language models effectively and thoroughly. *arXiv preprint arXiv:2505.10610*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Junjie Wu, Gefei Gu, Yanan Zheng, Dit-Yan Yeung, and Arman Cohan. 2025. Ref-long: Benchmarking the long-context referencing capability of long-context language models. *arXiv preprint arXiv:2507.09506*.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *The Thirteenth International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. 2025. Longproc: Benchmarking long-context language models on long procedural generation. *arXiv preprint arXiv:2501.05414*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and 1 others. 2024a. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and 1 others. 2024b. Infinitebench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277.

A Detailes of LVLMs

We list the detailed information of the models evaluated in the paper is presented in Table 6.

B Evaluation Metrics

To holistically assess the model performance, we assess both the quality of the citations and the factual correctness of the generated content. Our evaluation methodology leverages a powerful Multimodal Large Language Model (MLLM), GPT-4.1, as an automated judge for all metrics.

B.1 Citation Quality

We first evaluate the quality of citations in the generated responses using three metrics: Citation Recall (CR), Citation Precision (CP), and their harmonic mean, Citation F1. For each generated statement, the judge performs a natural language inference (NLI) style task to determine if the cited sources provide factual support. The sources in our multimodal context can be either text segments or images.

Citation Recall (CR) This metric measures whether the complete set of cited sources for a given statement provides sufficient factual support. For a response containing a set of statements S, we gather all cited sources $C(s)$ for each statement $s \in S$. The recall is then calculated as the fraction of statements that are fully supported by their cited sources.

$$\mathtt{CR} = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(f_{\text{eval}}(C(s), s) = \text{relevant}) \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function and f_{eval} is the MLLM-based evaluator. A score of 1 indicates that the collection of sources $C(s)$ factually entails the statement s.

Citation Precision (CP) This metric assesses the necessity of each individual citation, penalizing superfluous or irrelevant sources. We evaluate precision using a leave-one-out approach. A citation is deemed necessary only if the full set of sources supports the statement, but the set with this specific citation removed does not.

$$\mathtt{CP} = \frac{\sum_{s \in S} \sum_{c \in C(s)} \mathbb{I}_{\text{necessary}}(c, s)}{|\mathcal{C}_{\text{total}}|} \quad (3)$$

where $\mathcal{C}_{\text{total}}$ is the total number of citations in the response. The indicator $\mathbb{I}_{\text{necessary}}(c, s)$ is 1 if and only if both of the following conditions are met:

- The complete set of sources supports the statement: $f_{\text{eval}}(C(s), s) = \text{entailment}$.
- The set of sources without the specific citation c fails to support the statement: $f_{\text{eval}}(C(s) \setminus \{c\}, s) \neq \text{entailment}$.

Citation F1 (F1) We compute the F1-score as the harmonic mean of Citation Recall and Citation Precision, providing a single, comprehensive measure of overall citation quality.

$$\mathtt{F1} = \frac{2 \times \mathtt{CR} \times \mathtt{CP}}{\mathtt{CR} + \mathtt{CP}} \quad (4)$$

B.2 Generation Quality

Correctness (Cor) Beyond citation quality, we evaluate the factual correctness of the generated response by comparing it against a human-written ground-truth answer. The MLLM judge assigns a score on a 3-point scale: {0, 0.5, 1}. The criteria are as follows:

- **1 (Fully Correct):** The generated response accurately and completely covers the information present in the ground-truth answer. It contains no factual errors or significant omissions.
- **0.5 (Partially Correct):** The response is broadly correct but suffers from minor factual inaccuracies, hallucinations, or omits key information found in the ground-truth answer.
- **0 (Incorrect):** The response contains major factual errors, is irrelevant to the question, or is nonsensical.

The final correctness score is the average score over all instances in the test set. For a dataset with N instances, the score is calculated as:

$$\mathtt{Cor} = \frac{1}{N} \sum_{i=1}^{N} \text{Score}(R_i, A_i) \quad (5)$$

where R_i is the model-generated response for the i-th instance, A_i is the corresponding ground-truth answer, and the $\text{Score}(\cdot, \cdot)$ function represents the score assigned by the MLLM judge.

Model Name	Model Version	Context window	Model Link
Gemini-2.5-Pro	gemini-2.5-pro-preview-06-05	1000,000 tokens	-
Doubao-Seed-1.6	doubao-seed-1-6-250615	256,000 tokens	-
Qwen2.5-VL-3B	Qwen2.5-VL-3B-Instruct	128,000 tokens	https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct
Qwen2.5-VL-7B	Qwen2.5-VL-7B-Instruct	128,000 tokens	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
Qwen2.5-VL-32B	Qwen2.5-VL-32B-Instruct	128,000 tokens	https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct
Qwen2.5-VL-72B	Qwen2.5-VL-72B-Instruct	128,000 tokens	https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct
MiMo-VL-7B-SFT	MiMo-VL-7B-SFT-2508	128,000 tokens	https://huggingface.co/XiaomiMiMo/MiMo-VL-7B-SFT-2508
MiMo-VL-7B-RL	MiMo-VL-7B-RL-2508	128,000 tokens	https://huggingface.co/XiaomiMiMo/MiMo-VL-7B-RL-2508
Gemma3-12B	gemma-3-12b-it	131,072 tokens	https://huggingface.co/google/gemma-3-12b-it
Kimi-VL-A3B-Instruct	Kimi-VL-A3B-Instruct	131,072 tokens	https://huggingface.co/moonshotai/Kimi-VL-A3B-Instruct
Kimi-VL-A3B-Think	Kimi-VL-A3B-Thinking	131,072 tokens	https://huggingface.co/moonshotai/Kimi-VL-A3B-Thinking
GLM-4.5V	GLM-4.5V	65,536 tokens	https://huggingface.co/zai-org/GLM-4.5V

Table 6: Detailed information of models evaluated in MMLongCite.

You will receive a user's question about uploaded content (which could include text, images, video frames, etc.), a factual statement from an AI assistant's response based on that content, and part of the full content (since the full content is too long to display entirely). Your task is to carefully assess whether the provided content contains all the information of the statement. Please use the following grades to generate the rating:
- [[Relevant]] - All points of the statement are supported by the provided content.
- [[Irrelevant]] - The statement is unrelated to the provided content, or any points of the statement are inconsistent with the provided content.
Ensure that you do not use any information or knowledge outside of the provided content when evaluating.

Figure 4: Prompt for evaluating citation quality with GPT-4.1.

You are asked to evaluate the quality of the AI assistant's answer to user questions as an impartial judge, and your evaluation should take into account factors including correctness (high priority), and comprehensiveness (whether the assistant's answer covers all points).
Read the AI assistant's answer and compare against the reference answer, and give an overall integer rating in 1, 2, 3 (1 = wrong or irrelevant, 2 = partially correct, 3 = correct and comprehensive) based on the above principles.

Figure 5: Prompt for evaluating correctness with GPT-4.1.

Models	LongDocURL				MMLongBench-Doc				HotpotQA				2WikiMultiHopQA			
	CP	CR	F_1	Cor	CP	CR	F_1	Cor	CP	CR	F_1	Cor	CP	CR	F_1	Cor
🔒 *Proprietary LVLCMs*																
Gemini-2.5-Pro	91.29	89.66	89.93	82.74	91.69	91.64	90.92	72.00	96.06	98.09	96.63	94.58	96.39	95.92	95.83	79.44
Doubao-Seed-1.6(w/o think)	78.67	75.10	75.21	70.83	82.50	82.62	82.38	68.67	79.70	91.56	82.68	92.08	76.34	75.11	73.52	69.17
🔓 *Open-source LVLCMs*																
Qwen2.5-VL-3B	13.57	15.53	13.31	45.12	28.78	34.44	29.52	48.67	15.00	17.67	15.18	59.17	7.38	7.87	7.17	24.86
Qwen2.5-VL-7B	47.33	49.45	45.50	60.24	63.33	65.26	61.76	58.00	40.28	47.51	42.09	72.22	30.90	28.82	28.78	35.00
MiMo-VL-7B-SFT	68.15	45.77	51.17	68.69	69.67	55.33	58.77	65.00	72.12	68.65	68.29	78.89	68.50	56.99	59.94	50.42
MiMo-VL-7B-RL	79.16	43.56	52.50	75.48	80.89	57.78	63.81	67.67	75.40	70.10	69.88	84.58	70.75	54.43	58.76	51.25
Gemma3-12B	60.30	62.41	60.78	52.14	64.53	65.25	64.39	46.33	71.18	75.25	72.56	79.17	55.50	57.04	55.99	40.97
Kimi-VL-A3B-Instruct	62.21	57.06	56.63	58.93	62.25	57.80	57.41	64.33	61.67	55.33	56.79	73.06	39.29	31.77	34.21	37.36
Kimi-VL-A3B-Think	42.07	20.93	26.34	50.36	65.88	38.70	45.88	63.33	43.45	23.27	27.96	58.61	32.56	15.26	19.29	40.14
Qwen2.5-VL-32B	63.04	45.56	50.59	68.45	78.77	58.21	64.43	63.67	57.82	50.32	51.73	81.67	48.44	38.06	41.23	50.56
Qwen2.5-VL-72B	79.67	77.12	77.12	70.60	83.06	80.89	80.81	65.33	64.40	67.80	64.95	83.47	48.75	48.47	47.78	47.36
GLM-4.5V(w/o think)	6.36	65.82	67.78	73.33	80.72	72.13	74.53	63.33	73.92	82.27	75.41	88.61	63.02	58.67	58.57	59.58
GLM-4.5V(w/ think)	80.53	65.48	68.91	79.40	80.37	68.99	71.42	70.33	76.60	79.68	74.75	92.64	76.41	68.12	68.26	73.75

Table 7: Performance of LVLMs in Single-Source Visual Reasoning and Multi-Source Visual Reasoning of MMLongCite.

Models	Visual Haystack				MM-NIAH				Video-MME				LongVideoBench			
	CP	**CR**	**F_1**	**Cor**	**CP**	**CR**	**F_1**	**Cor**	**CP**	**CR**	**F_1**	**Cor**	**CP**	**CR**	**F_1**	**Cor**
🔒 ***Proprietary LVLCMs***																
Gemini-2.5-Pro	88.69	92.48	90.04	87.27	89.02	88.57	88.07	84.44	83.47	70.58	74.16	79.00	82.57	83.43	81.52	68.89
Doubao-Seed-1.6(w/o think)	76.46	71.46	72.96	76.21	85.09	83.80	84.10	86.25	72.25	65.70	65.59	75.50	71.36	70.67	68.42	59.44
🔓 ***Open-source LVLCMs***																
Qwen2.5-VL-3B	11.21	10.71	10.47	25.30	14.38	9.41	10.90	70.28	34.83	34.33	32.31	66.50	40.09	46.02	40.62	56.67
Qwen2.5-VL-7B	26.96	19.16	21.30	51.21	17.52	17.05	16.75	62.22	52.83	44.20	45.37	68.00	51.30	45.35	44.72	63.89
MiMo-VL-7B-SFT	53.28	42.11	44.69	63.18	46.08	36.90	39.20	65.83	63.55	27.93	36.39	70.00	65.09	31.85	39.82	70.00
MiMo-VL-7B-RL	71.41	42.29	49.47	66.44	65.32	48.34	53.12	72.29	65.75	22.98	32.23	70.50	69.88	27.47	36.64	72.78
Gemma3-12B	49.10	46.99	47.21	44.32	33.38	31.61	32.07	73.75	63.93	72.92	67.30	72.00	72.15	79.85	75.04	55.00
Kimi-VL-A3B-Instruct	23.26	19.20	20.45	52.80	20.43	20.37	20.29	72.29	7.33	5.37	5.89	61.50	7.87	6.39	6.66	61.11
Kimi-VL-A3B-Think	43.59	24.70	30.03	46.82	19.73	10.65	12.64	37.50	27.69	9.91	13.44	48.50	58.03	23.05	30.90	62.22
Qwen2.5-VL-32B	44.30	20.82	27.65	55.23	66.68	52.87	57.42	82.64	62.72	27.52	35.76	73.00	65.33	41.32	47.91	73.89
Qwen2.5-VL-72B	63.74	52.80	56.31	60.61	66.27	64.27	64.86	85.97	69.93	59.08	61.43	74.00	77.50	61.20	65.58	63.89
GLM-4.5V(w/o think)	62.46	50.37	54.23	70.45	49.91	42.74	44.19	83.12	67.81	39.92	47.25	78.50	73.98	52.22	58.21	68.89
GLM-4.5V(w/ think)	68.25	37.19	45.70	75.38	53.04	43.14	45.76	74.51	65.11	32.84	41.63	80.50	62.46	33.30	40.70	72.22

Table 8: Performance of LVLMs in Vision Grounding and Video Understanding of MMLongCite.