MMLDSum-LLM: Multimodal Long-Document Summarization with Visual-Alignment and Keyword-Aware

Xianpeng Zhang and Jiahua Yang and Dongyu Chen and Lei Zhang and Jian Ma
Xu Guohuan and Haonan Lu and Tianhuang Su and Chuangchuang Wang and Kai Tang
OPPO Guangdong Mobile Telecommunications Co., Ltd.

Abstract

Multimodal long documents are core carriers of professional knowledge, where critical evidence is sparsely distributed across paragraphs and modalities. This easily causes key information omission and cross-modal hallucinations in summarization by multimodal LLMs. These issues stem from attention drift in long-range dependency modeling and gaps in inter-modal alignment. To address this, we introduce **MMLDSum-Bench**, a high-quality benchmark for multimodal long-document summarization, covering multiple domains, context-length scales, and visual-textual modality distributions. We further propose **MMLDSum-LLM**, a reproducible two-stage training framework that combines supervised fine-tuning with *visual-alignment weighted loss* and *keyword-aware weighted loss*, followed by GRPO with a multi-objective reward (keyword coverage, image-text alignment, ROUGE, and length control). Extensive experiments on MMLDSum-Bench evaluate our approach against leading closed-source and open-source multimodal models under a unified protocol that incorporates LLM-as-a-judge scoring, atomic-claim precision/recall, image-text alignment (ITA), and ROUGE. The results demonstrate that our approach significantly improves key-information coverage and cross-modal consistency.

1 Introduction

In an era of information explosion (Goyal et al., 2022), multimodal long documents, such as academic papers, medical reports, and financial annual reports, have become the dominant medium for knowledge transmission in professional domains. Such documents integrate multiple modalities, including text, figures, and tables, with each contributing distinct yet complementary information. Crucially, these modalities do not function in isolation but mutually reinforce and corroborate one another, collectively delivering the full informational content of the document. Multimodal long-document summarization aims to condense such a document into a concise, coherent natural-language summary that faithfully captures the salient information across all modalities, preserves cross-modal evidential consistency, and retains the key factual relations between textual arguments and their supporting visual elements.

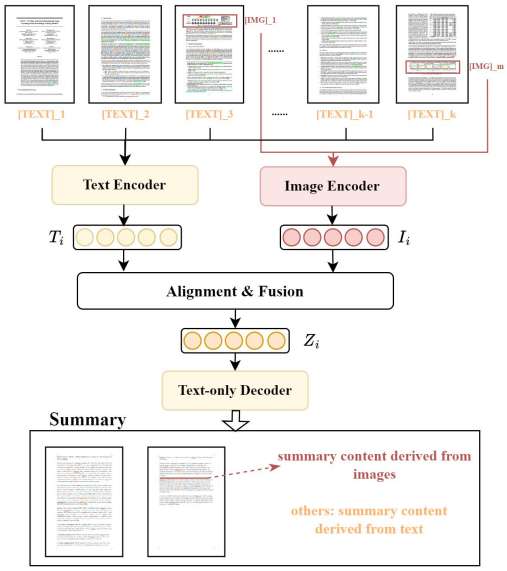


Figure 1: Illustration of conventional multimodal long-document summarization.

Early summarization studies focused on extractive methods (e.g., TF-IDF, TextRank) and later shifted to neural abstractive models (Rush et al., 2015). With the rise of large language models, summarization has benefited from stronger generation quality and controllable prompting. In parallel, multimodal summarization extends beyond text by incorporating images and other modalities, often requiring explicit cross-modal alignment to avoid modality suppression and hallucinations (Jangra et al., 2023). However, most existing multimodal summarization benchmarks and methods primar-

ily focus on short contexts or domain-specific settings (e.g., dialogue/video), and do not capture the sparse, cross-modally dispersed distribution of key evidence in multimodal long documents (Kumbhar et al., 2023; Khilji et al., 2023).

At present, multimodal long-document summarization faces severe challenges at both the data and methodological levels (Koh et al., 2022). **On the data side**, most existing multimodal summarization datasets are confined to specific domains — such as dialogues, news, and clinical reports — or limited to short contexts, leaving long multimodal documents with diverse visual-textual modality distributions substantially underrepresented. Although several long-context multimodal benchmarks have recently emerged, summarization-specific supervision and evaluation protocols under long-context settings remain scarce (Wang et al., 2025). **On the method side**, models are required to jointly address long-range dependency modeling, cross-modal grounding, and information selection. Failures in these aspects typically manifest as both missing key evidence and cross-modal hallucinations. As illustrated in Figure 1, conventional multimodal long-document summarization approaches generally adopt a pipeline architecture consisting of modality-specific encoding, feature alignment and fusion, and decoder-based generation. Textual and visual features are first extracted independently by text and image encoders, aligned and fused into a joint representation, and subsequently decoded to produce a textual summary. Some studies further select images that are most semantically relevant to the generated summary to yield multimodal summary outputs. However, such paradigms remain fundamentally limited by constrained long-sequence modeling capacity, unstable cross-modal semantic alignment, and inadequate mechanisms for effective information selection (Hua et al., 2025).

These failures stem from three intertwined challenges: (i) attention drift over long sequences causes models to over-attend to local context and miss globally salient evidence (Ouyang et al., 2022); (ii) cross-modal misalignment causes visually distant evidence to be suppressed or hallucinated; and (iii) standard SFT objectives treat all tokens uniformly, giving insufficient weight to sparse but critical evidence spans.

To address these challenges, we propose MMLDSum-LLM (Multimodal Long-Document Summarization with Visual-Alignment and Keyword-Aware Training), a two-stage framework integrating supervised fine-tuning (SFT) and group relative policy optimization (GRPO) (Shao et al., 2024). Motivated by a cognitive anchoring strategy, in which readers first anchor core concepts and salient visuals before organizing supporting details, we design a composite weighted SFT loss with two complementary components: a visual-alignment weight that amplifies learning on image-associated spans, and a keyword-aware weight that emphasizes TF-IDF-filtered key entities. GRPO then optimizes sequence-level objectives via multi-objective verifiable rewards for keyword coverage, image-text alignment, ROUGE, and length control. We also introduce MMLDSum-Bench, a benchmark covering six domains, five context-length scales (4k–64k tokens), and four visual-textual modality distribution categories, providing a comprehensive testbed for this task.

Our contributions are summarized as follows:

- We construct **MMLDSum-Bench**, a high-quality benchmark for multimodal long-document summarization, providing a comprehensive and realistic testbed for this task.
- We design a systematic **evaluation protocol** encompassing LLM-as-a-judge scoring, atomic-claim precision/recall/F1, image-text alignment (ITA), and ROUGE, and conduct a unified comparative evaluation of state-of-the-art closed-source and open-source multimodal models on MMLDSum-Bench.
- We propose **MMLDSum-LLM**, a two-stage training framework that combines visual-alignment and keyword-aware weighted SFT with GRPO-based reinforcement learning using multi-objective verifiable rewards. Experiments demonstrate that MMLDSum-LLM significantly improves key-information coverage and cross-modal consistency.

2 Related Work

2.1 Multimodal Summarization

LLMs have substantially improved text summarization in generation quality and instruction following (Narayan et al., 2021; Adams et al., 2023). Multimodal summarization extends this by incorporating images and other modalities via modality-specific encoders, cross-modal fusion, and contrastive alignment (Li et al., 2018; He et al., 2023),

with retrieval-augmented methods further improving visual grounding (Rafi and Das, 2024). Research spans domain-specific settings including medical imaging (Ghosh et al., 2024; Lu et al., 2022) and dynamic scenarios such as dialogue and video summarization (Lu et al., 2024; Qiu et al., 2024; Hua et al., 2025; Yang et al., 2024; Mahon and Lapata, 2024; Tan et al., 2025).

2.2 Long-Context Vision–Language Models

Long-Context VLMs (LCVLMs) (Song et al., 2025) enable end-to-end multimodal understanding at scale (Wang et al., 2025), but remain limited for long-document summarization: their alignment modules are designed for shorter sequences, causing semantic drift when evidence is asynchronously distributed across long documents (Bai et al., 2024; Wan et al., 2025), and pre-training objectives target general understanding rather than the selective compression required for quality summaries (Deng et al., 2025). MMLDSum-LLM directly addresses these gaps through explicit visual-alignment weighting and keyword-aware supervised training.

3 MMLDSum-Bench

The MMLDSum-Bench benchmark targets the **multimodal long-document summarization** task: given the textual content of a document and its associated image set, the model is required to generate a natural-language summary under a length constraint that captures core factual information and critical visual evidence while preserving cross-modal consistency. The benchmark contains approximately 5k (5,149) multimodal long documents paired with over 40k associated images across diverse domains. We stratify documents into five context-length scales (4k–64k tokens, with an average length of $\sim$ 25k tokens) and categorize the data into four categories of visual-textual modality distributions, spanning the full spectrum from heavily text-dominant to heavily image-dominant settings.

As illustrated in Figure 2, the corpus covers multiple domains, including academia, medicine, finance, news, technology and others, enabling representative sampling of both narrative-heavy and evidence-heavy documents. In terms of context length, the dataset is concentrated in the 4k–16k range while also containing a substantial number of samples in the 16k–64k regime, which is sufficient for evaluating long-context summarization performance of multimodal large models. Table 1 further shows broad coverage across modality distributions. Although the benchmark is dominated by heavily text-dominant documents, accounting for 80.9% (4,163 samples), it also includes meaningful proportions of lightly text-dominant (11.9%, 611 samples) and lightly image-dominant (6.6%, 342 samples) documents. These distribution characteristics ensure that the benchmark provides comprehensive coverage across three key dimensions (**domain, context length, and visual-textual modality distribution**) rather than a single narrow regime, thereby establishing a realistic and reliable data environment for multimodal long-document summarization research.

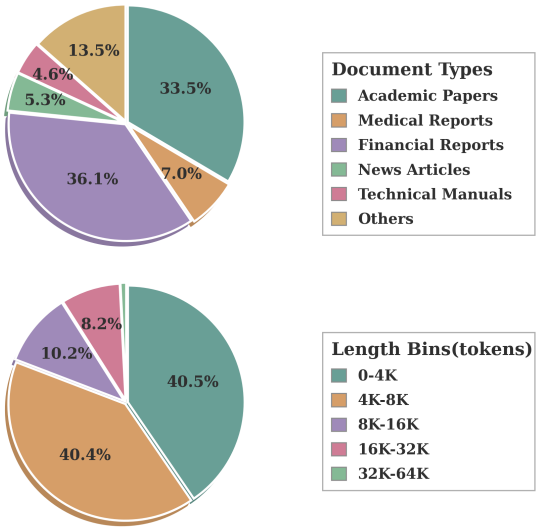


Figure 2: Domain distribution of MMLDSum-Bench.

Table 1: Distribution of Image–Text Ratios in the Dataset

Type	Image Ratio	Count	Percentage (%)
Heavily Text-Dominant	0–0.25	4163	80.9
Lightly Text-Dominant	0.25–0.5	611	11.9
Lightly Image-Dominant	0.5–0.75	342	6.6
Heavily Image-Dominant	0.75–1	33	0.6

As shown in Figure 5, we employ a three-stage pipeline to balance quality and cost:

(i) **Data Processing**: This stage performs document chunking and global signal extraction. Each document is segmented by paragraph boundaries under a length threshold (approximately 3k tokens), with images assigned to chunks according to their original positions or adjacent paragraphs. Doubao-1.5-pro-256k is then used to extract global signals,

including topics, outlines, and key entities.

(ii) **Multimodal Summary Generation**: Gemini-2.5-Pro first produces local summaries for individual chunks. These local summaries are then fused with the extracted global signals (topics, outlines, key entities) to generate candidate global summaries.

(iii) **Quality Verification and Regeneration**: Candidate summaries are evaluated through a multi-model scoring-and-voting mechanism (GPT-4o, Doubao-seed-1.6, Gemini-2.5-Pro) across five dimensions: completeness, accuracy, coherence, conciseness, and overall quality. A candidate is accepted only when all three models assign scores above a predefined threshold; otherwise, the summary generation process is re-executed until the consensus criterion is met.

To ensure robust dataset quality assessment, we conduct stratified human evaluation on 600 summaries across domain, context-length scale, and visual-textual modality distribution. Each summary is evaluated along five dimensions (completeness, accuracy, coherence, conciseness, and overall quality), and we report per-dimension mean scores with 95% confidence intervals, together with per-dimension inter-annotator agreement. We further perform claim-level manual verification on 200 atomic claims to directly assess factual correctness and evidence-grounding consistency. In addition, we quantify the contribution of the regeneration module by reporting before/after quality statistics for regenerated samples. Detailed protocols and full results are provided in Appendix A.3. Overall, the evaluation indicates high annotation quality (overall mean score: 4.7/5.0; overall Cohen's Kappa: 0.83).

4 Methodology

4.1 Task Definition

Let a multimodal document be $x = (T, I)$, where T is the text token sequence and I is the image set. Given x, the model generates a summary $y = (y_1, \ldots, y_n)$ with conditional distribution $p_\theta(y \mid x)$. As discussed in Section 1, multimodal long-document summarization mainly suffers from two issues: (i) omission of key information caused by attention drift over long contexts, and (ii) cross-modal hallucination caused by text–image misalignment. Therefore, our goal is not only to maximize conditional likelihood, but also to improve factual/visual evidence coverage and cross-modal consistency under a length budget.

As shown in Figure 3, we optimize this goal with a two-stage framework. Stage 1 (anchor-weighted SFT) identifies textual and visual anchors and increases supervision on anchor-related spans to strengthen local grounding. Stage 2 (GRPO-based RL) optimizes sequence-level quality, including key-information coverage, cross-modal consistency, and conciseness. This local-to-global optimization forms the core of MMLDSum-LLM.

4.2 Stage 1: Visual-Alignment and Keyword-Aware Weighted SFT

Limitation of standard cross-entropy. Given training pairs $(x, y^\star)$, the standard token-level cross-entropy objective is:

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{t=1}^{|y^\star|} \log p_\theta(y_t^\star \mid y_{<t}^\star, x). \quad (1)$$

This objective assigns equal importance to all reference tokens, which weakens supervision on sparse but critical evidence tokens. As a result, the model may miss key facts or generate visually unsupported content. We therefore introduce a weighted strategy to strengthen learning on evidence-critical positions.

Following the cognitive anchoring principle, we first amplify learning signals on visually grounded spans. During data construction, summary spans that describe or reference visual evidence are marked via special-token matching and regular-expression rules. We define an indicator $\mathbb{I}_t^{\mathrm{img}} \in \{0, 1\}$ that equals 1 if token $y_t^\star$ belongs to a visually grounded span, and apply a per-token weight:

$$w_t^{\mathrm{img}} = 1 + \lambda_{\mathrm{img}} \cdot \mathbb{I}_t^{\mathrm{img}}. \quad (2)$$

In parallel, we build a keyword set K as textual fact anchors. We extract subject–verb–object (SVO) tuples with a dependency parser, then apply TF-IDF filtering to keep domain-salient entities and relations. Let $\mathbb{I}_t^{\mathrm{kw}} \in \{0, 1\}$ indicate whether token $y_t^\star$ matches an extracted keyword:

$$w_t^{\mathrm{kw}} = 1 + \lambda_{\mathrm{kw}} \cdot \mathbb{I}_t^{\mathrm{kw}}. \quad (3)$$

The final SFT loss fuses both weights additively to amplify learning signals on visual evidence and key facts:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{t=1}^{|y^\star|} \left(w_t^{\mathrm{img}} + w_t^{\mathrm{kw}} \right) \log p_\theta(y_t^\star \mid y_{<t}^\star, x). \quad (4)$$

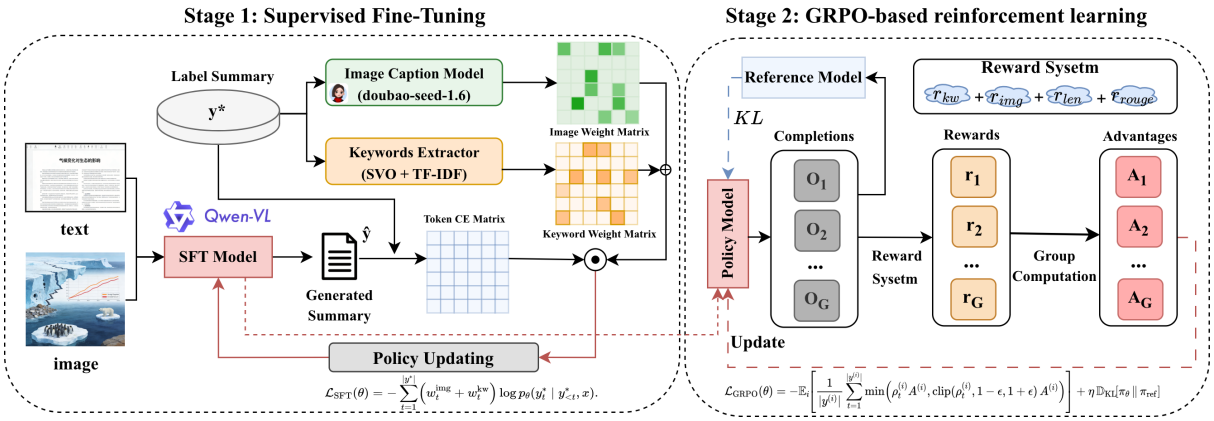


Figure 3: Overview of MMLDSum-LLM: alignment signal acquisition, weighted SFT, and GRPO-based reinforcement learning with multi-objective verifiable rewards.

We use additive fusion so each signal contributes independently: tokens matched by either type are still reinforced, unlike multiplicative fusion, which mainly boosts rare co-occurrences. Hyperparameters λ_{img} and λ_{kw} control weighting strength; values in the range 5–7 provide a good balance between evidence coverage and fluency.

4.3 Stage 2: GRPO-Based Reinforcement Learning

Visual-alignment and keyword-aware weighted SFT strengthens token-level supervision on key evidence, but it is still imitation learning and remains tied to the training distribution. It also cannot directly optimize summary-level properties—key-information coverage, cross-modal consistency, and conciseness. To address this, we add a second stage using GRPO (Shao et al., 2024), which evaluates each sample against the within-group mean of G candidate summaries. We use a composite reward with four components:

$$r(y; x) = \alpha\, r_{\text{kw}} + \beta\, r_{\text{img}} + \gamma\, r_{\text{rouge}} + \delta\, r_{\text{len}}. \quad (5)$$

- r_{kw} (*keyword coverage*): mitigates key-information omission by measuring precision, recall, and F1 between generated-summary keywords and source fact anchors.
- r_{img} (*image-text alignment*): mitigates cross-modal hallucination by computing semantic similarity between summary segments and image captions from an auxiliary captioning model.
- r_{rouge} (*ROUGE score*): uses the average of ROUGE-1/2/L against the reference summary as a general quality signal.
- r_{len} (*length control*): discourages overly long outputs and controls RL-induced length inflation.

All four rewards are rule-based and deterministic, casting training as reinforcement learning with verifiable rewards (RLVR) and avoiding costly, unstable LLM-based reward models. We set $\alpha = 0.5$, $\beta = 0.2$, $\gamma = 0.15$, and $\delta = 0.15$, prioritizing keyword coverage because key-information omission is the dominant failure mode in preliminary experiments.

For each input x, we sample a group of G candidate summaries $\{y^{(i)}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ from the policy snapshot and score each one with the composite reward $r^{(i)}$ defined in Eq. 5. Following GRPO (Shao et al., 2024), we standardize rewards within the group:

$$A^{(i)} = \frac{r^{(i)} - \text{mean}\left(\{r^{(j)}\}_{j=1}^{G}\right)}{\text{std}\left(\{r^{(j)}\}_{j=1}^{G}\right) + \varepsilon}, \quad (6)$$

The policy is then updated with a token-level clipped objective regularized toward a fixed reference policy π_{ref}, which we initialize from the Stage 1 SFT checkpoint and keep frozen throughout RL:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\,\mathbb{E}_i\left[\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\min\Big(\rho_t^{(i)} A^{(i)},\ \text{clip}(\rho_t^{(i)},\, 1-\epsilon,\, 1+\epsilon)\, A^{(i)}\Big)\right] + \eta\, \mathbb{D}_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}]\,, \quad (7)$$

where the per-token importance ratio is

$$\rho_t^{(i)} = \frac{\pi_\theta(y_t^{(i)} \mid y_{<t}^{(i)}, x)}{\pi_{\theta_{\text{old}}}(y_t^{(i)} \mid y_{<t}^{(i)}, x)}, \quad (8)$$

The KL term preserves Stage 1 priors (visual alignment and keyword grounding) while allowing stable optimization of summary-level rewards.

The two stages are complementary: Stage 1 improves local evidence grounding through token weighting, and Stage 2 improves global summary quality and generalization through reward-driven exploration.

5 Experiments

5.1 Evaluation Metrics

To comprehensively evaluate multimodal long-document summarization, we build a multidimensional automatic evaluation suite (Liu et al., 2025; Langston and Ashford, 2024) with four complementary metric families. Each family focuses on a different quality dimension, and their combination enables cross-validation over semantic fidelity, cross-modal consistency, and surface-level text quality. If a model shows stable gains across all metrics, this provides strong evidence of substantive summary quality improvement.

- **LLM-as-a-judge**: We use both GPT-4o and GPT-5 as judges to improve scoring credibility and enable cross-judge consistency. They score each summary on completeness, accuracy, coherence, conciseness, and overall quality, with three runs per sample averaged to reduce variance.

- **Atomic-claim precision/recall**: GPT-4o extracts atomic factual claims from reference and generated summaries, and computes precision, recall, and F1 via semantic matching. Compared with holistic judge scores, this metric offers finer-grained measurement of factuality (precision) and completeness (recall), and does not require access to full source documents at evaluation time (Zhang et al., 2025).

- **Image-Text Alignment**: We generate captions for document images and compute semantic similarity between summary segments and captions using BGE-M3 (Chen et al., 2024) (threshold 0.65), then report recall (Hua et al., 2025). ITA measures whether key visual evidence is faithfully reflected in the summary.

- **ROUGE**: ROUGE-1, ROUGE-2, and ROUGE-L measure n-gram overlap with the reference summary, providing a lightweight indicator of coverage and surface text quality.

Image-Text Alignment and ROUGE are also used as reward components in the GRPO stage (Section 4.3). To ensure gains come from real quality improvement rather than reward fitting, we treat LLM-as-a-judge scores and atomic-claim precision/recall as independent validation metrics and exclude them from training objectives. When improvements in ITA and ROUGE are accompanied by stable gains in judge scores and atomic-claim metrics, this jointly verifies genuine multidimensional quality improvement rather than metric gaming.

5.2 Experimental Setup

Baselines. We conduct comparative experiments on the MMLDSum-Bench benchmark, covering representative closed-source and open-source multimodal models, and build backbone-matched baselines to ensure fair comparison. The closed-source group includes GPT-5, Claude-4-Sonnet, Doubao-Seed-1.6, Qwen-VL-Max, Qwen3-VL-Plus, and Step-1o-Vision-32k. The open-source group includes strong community baselines across different scales and architectures: Qwen2.5-VL, Qwen3-VL, InternVL3.5, Gemma3, and Phi-4-Multimodal-Instruct, spanning lightweight to large-parameter settings for different deployment scenarios. To avoid evaluation bias, Gemini-2.5-Pro and GPT-4o are excluded, since they are already used in our data construction and evaluation pipeline (Section 3 and Section 5.1). To verify the effectiveness of our two-stage training framework, we build SFT-only baselines on open-source backbones, including Qwen2.5-VL (3B/7B) and Qwen3-VL (8B), and compare them directly with MMLDSum-LLM. All models are evaluated under identical settings: the same test split, length-control strategy, prompt template, and a unified automated evaluation script for all metrics, ensuring fair and comparable results.

Prompting and decoding. For all models, we use a unified instruction template that (i) asks for a concise global summary, (ii) explicitly requests grounding to both text and figures, and (iii) constrains output length. For fair comparison, we enforce the same maximum output token budget and use deterministic decoding (temperature $= 0$) unless a model requires sampling.

Implementation details. For SFT, we train the model for 3 epochs using the AdamW optimizer with a learning rate of 5×10^{-6} and a batch size of 1. For GRPO, we employ a group size of $G =$

Table 2: Comparison results on MMLDSum-Bench across closed-source models, open-source models, other methods, and our MMLDSum-LLM variants. Bold numbers denote the best-performing metrics.

Model	Max ctx	GPT-4o score					GPT-5 score					Atomic claim		ITA	ROUGE		
		Comp.	Acc.	Conc.	Coh.	Overall	Comp.	Acc.	Conc.	Coh.	Overall	R	F1	ITA-R	R-1	R-2	R-L
Closed-source models																	
step-1o-vision-32k	32k	4.37	4.88	4.95	4.94	4.57	3.19	**4.28**	4.64	4.82	3.51	0.47	0.60	0.49	0.42	0.20	0.25
claude-4-sonnet	1000k	4.34	4.91	4.94	4.95	4.52	3.48	3.79	**4.67**	4.60	3.58	0.67	0.75	0.59	0.50	0.24	0.30
qwen-vl-max	128k	4.57	4.94	4.95	4.98	4.75	4.00	3.64	4.16	4.85	3.70	0.66	0.73	**0.72**	0.52	0.23	0.29
qwen3-vl-plus	256k	**4.66**	**4.96**	**4.96**	**4.99**	**4.83**	**4.08**	3.79	4.07	4.88	3.76	0.71	0.77	0.71	**0.55**	0.24	0.30
doubao-seed-1.6	256k	4.58	4.94	**4.96**	4.98	4.77	3.98	4.06	4.49	**4.90**	**3.87**	0.71	0.77	0.66	**0.55**	**0.27**	**0.34**
gpt-5	128k	4.64	4.94	4.86	4.96	4.79	–	–	–	–	–	**0.90**	**0.85**	**0.72**	0.53	0.20	0.30
Open-source models																	
phi-4-multimodal-instruct	128k	1.87	1.82	2.28	2.10	1.76	1.25	1.31	1.56	1.82	1.28	0.16	0.18	0.29	0.09	0.02	0.06
qwen2.5-vl-32b-instruct	128k	3.78	4.51	4.33	4.61	4.04	2.30	2.45	2.47	3.34	2.37	0.42	0.52	0.63	0.33	0.10	0.16
qwen2.5-vl-72b-instruct	128k	3.67	4.42	4.11	4.46	3.90	2.24	2.43	2.32	3.09	2.24	0.43	0.51	0.65	0.25	0.07	0.12
internvl3.5-14b-instruct	32k	4.11	4.76	4.81	4.83	4.36	2.86	3.53	4.36	4.51	3.11	0.47	0.58	0.54	0.35	0.15	0.20
internvl3.5-38b-instruct	32k	4.01	4.74	4.76	4.80	4.29	2.81	3.63	4.34	4.45	3.12	0.42	0.54	0.51	0.30	0.12	0.17
gemma3-12b	128k	4.20	4.81	4.89	4.90	4.43	2.96	3.39	4.49	4.56	3.17	0.46	0.58	0.59	0.36	0.16	0.21
gemma3-27b	128k	4.20	**4.86**	**4.93**	4.92	4.46	3.05	3.62	**4.61**	4.60	**3.31**	0.48	0.61	0.57	0.31	0.13	0.18
qwen3-vl-32b-instruct	256k	4.16	4.81	4.65	4.89	4.36	2.75	2.30	2.19	3.53	2.46	0.58	0.62	0.78	0.35	0.09	0.15
qwen3.5-vl-27b	128k	3.88	4.22	4.26	4.31	4.00	3.18	2.82	3.34	4.03	2.94	0.63	0.63	0.47	0.43	0.17	0.23
Other methods																	
qwen2.5-vl-7b-cod	128k	3.64	4.30	4.38	4.31	3.90	2.56	3.34	4.31	4.25	2.90	0.30	0.41	0.55	0.26	0.11	0.15
qwen3-vl-8b-cod	256k	4.20	4.71	4.56	4.73	4.45	3.71	2.83	3.99	4.63	3.25	0.63	0.69	0.60	0.37	0.14	0.21
longwriter-llama3.1-8b-caption	128k	3.58	4.31	4.48	4.34	3.87	2.42	3.76	4.47	4.39	2.88	0.29	0.40	0.41	0.18	0.07	0.11
longwriter-glm4-9b-caption	128k	3.42	4.30	4.36	4.25	3.72	2.53	**4.00**	3.58	3.88	2.80	0.37	0.44	0.60	0.24	0.08	0.12
Ours																	
qwen2.5-vl-3b-sft	128k	2.92	3.42	3.45	3.55	3.15	1.86	1.51	1.98	2.48	1.68	0.34	0.38	0.50	0.26	0.06	0.12
MMLDSum-qwen2.5-vl-3b	128k	3.48	4.18	4.08	4.37	3.79	2.26	1.82	2.43	3.10	2.00	0.53	0.54	0.76	0.38	0.12	0.17
qwen2.5-vl-7b-sft	128k	3.48	4.22	4.51	4.44	3.86	2.43	2.20	3.45	3.87	2.42	0.46	0.53	0.72	0.41	0.14	0.20
MMLDSum-qwen2.5-vl-7b	128k	3.82	4.65	4.60	4.79	4.13	2.63	2.47	3.66	4.01	2.58	0.54	0.59	0.87	0.51	0.21	0.26
qwen3-vl-8b-sft	256k	4.10	4.73	4.63	4.83	4.29	3.47	2.68	3.04	4.13	2.78	0.73	0.73	0.82	0.51	0.21	0.28
MMLDSum-qwen3-vl-8b	256k	**4.33**	4.85	4.78	**4.93**	**4.51**	**4.08**	3.76	3.65	**4.66**	3.21	**0.85**	**0.80**	**0.89**	**0.63**	**0.30**	**0.37**

5, a clipping parameter of $\epsilon = 0.01$, and a KL regularization coefficient of $\eta = 0.01$. The model is trained for 15 epochs with a learning rate of 10^{-6}. All experiments are conducted on 8 NVIDIA H20 GPUs.

5.3 Quantitative Results

Table 2 presents the comprehensive evaluation on MMLDSum-Bench across all four metric families: LLM-as-a-judge scoring from GPT-4o and GPT-5 (completeness, accuracy, conciseness, coherence, and overall), atomic-claim recall/F1, image-text alignment (ITA-R), and ROUGE. Overall, MMLDSum-LLM consistently improves key-information coverage and cross-modal consistency, with the largest gains on dimensions that directly reflect *completeness* (GPT-4o/GPT-5 completeness and atomic recall) and *visual evidence alignment* (ITA-R).

MMLDSum-qwen3vl-8b achieves open-source SOTA and approaches top closed-source models. MMLDSum-qwen3vl-8b reaches a GPT-4o overall score of 4.51 and a GPT-5 overall score of 3.21, surpassing all open-source baselines and approaching leading closed-source systems (Claude-4-Sonnet: 4.52/3.58; Step-1o-Vision-32k: 4.57/3.51). On atomic-claim recall—the direct signal of factual completeness—our model achieves 0.85, approaching GPT-5 (0.90) and substantially outperforming all other closed-source models (next best: Qwen3-VL-Plus and Doubao-Seed-1.6 at 0.71).

Two-stage training yields consistent gains across all backbone sizes. On Qwen3-VL-8B, the two-stage framework raises GPT-4o overall from 4.29 to 4.51 (+5.1%), GPT-5 overall from 2.78 to 3.21 (+15.5%), and atomic recall from 0.73 to 0.85 (+16.4%). On Qwen2.5-VL-7B, GPT-4o completeness improves by +9.8% (3.48 → 3.82), GPT-4o overall by +7.0% (3.86 → 4.13), and ITA-R by +20.8% (0.72 → 0.87); the 7B model surpasses Qwen3-VL-32B on ITA-R (0.87 vs. 0.78) with four times fewer parameters. Even on the 3B backbone, GPT-4o overall gains +20.3% (3.15 → 3.79) and ITA-R improves by +52.0% (0.50 → 0.76), exceeding Qwen2.5-VL-32B on ITA-R (0.76 vs. 0.63).

Closed-source models lead on judge scores, yet coverage gaps persist across all systems. Closed-source models achieve consistently high judge scores (GPT-4o overall: 4.52–4.83), but atomic-claim recall lags precision across most models—even GPT-5 (0.90 aggregate recall) degrades at 64k tokens (Appendix D)—confirming

Table 3: Ablation study of MMLDSum-LLM. The final model is highlighted and annotated with improvement over the SFT baseline.

Variant	GPT-4o score					GPT-5 score					Atomic claim			ITA	ROUGE		
	Comp.	Acc.	Conc.	Coh.	Overall	Comp.	Acc.	Conc.	Coh.	Overall	P	R	F1	ITA-R	R-1	R-2	R-L
qwen2.5-vl-7b-sft	3.48	4.22	4.51	4.44	3.86	2.43	2.20	3.45	3.87	2.42	0.69	0.46	0.53	0.72	0.41	0.14	0.20
qwen2.5-vl-7b-sft (image_weight, **I**)	3.52	4.25	4.48	4.43	3.87	2.45	2.26	3.42	3.89	2.41	0.71	0.51	0.56	0.73	0.40	0.14	0.19
qwen2.5-vl-7b-sft (keywords_weight,**K**)	3.55	4.31	4.54	4.52	3.91	2.51	2.24	3.56	4.01	2.46	0.73	0.51	0.57	0.74	0.44	0.16	0.22
qwen2.5-vl-7b-sft (**I+K**)	3.57	4.29	4.51	4.50	3.89	2.54	2.28	3.54	3.94	2.47	0.70	0.48	0.55	0.75	0.42	0.15	0.21
qwen2.5-vl-7b-sft + grpo	3.78	4.61	4.58	4.74	4.07	2.63	2.45	3.70	4.10	2.61	0.70	0.54	0.59	0.83	0.49	0.19	0.24
MMLDSum-qwen2.5-vl-7b(ours)	**3.82**	**4.65**	**4.60**	**4.79**	**4.13**	**2.63**	**2.47**	3.66	4.01	2.58	**0.71**	**0.54**	**0.59**	**0.87**	**0.51**	**0.21**	**0.26**
↑(%)	**9.77**	**10.19**	**2.00**	**7.88**	**6.99**	**8.23**	**12.27**	**6.09**	**3.62**	**6.61**	**2.90**	**17.39**	**11.32**	**20.83**	**24.39**	**50.00**	**30.00**

that fully faithful long-context summarization remains an open problem.

SFT-only baselines still exhibit omissions and cross-modal inconsistencies, reflecting the limits of token-level cross-entropy on sparse evidence. Both judges yield convergent rankings (Overall Spearman ρ=0.894, see Table 7 in Appendix D;GPT-5 applies a stricter standard); boundary cases include *Phi-4-Multimodal-Instruct* (1.76, limited Chinese capability) and *Step-1o-Vision-32k* (32k context ceiling). Length-stratified heatmaps (Appendix D, Figures 7–11) confirm that MMLDSum-LLM's advantage is most pronounced in the 16k–64k range, where it achieves the best trade-off among open-source models across all four metric families.

5.4 Ablation Study

Table 3 validates the contribution of each component on the Qwen2.5-VL-7B backbone. Visual-alignment weighting (**I**) primarily boosts cross-modal consistency (ITA-R: $0.72 \rightarrow 0.73$, $+1.4\%$), while keyword-aware weighting (**K**) primarily improves key-fact retention (atomic recall: $0.46 \rightarrow 0.51$, $+10.9\%$). Effects are not isolated: **K** also lifts ITA-R to 0.74, and **I** also benefits atomic precision. Combining both (**I+K**) further raises ITA-R to 0.75. Adding GRPO yields substantially larger sequence-level gains—ITA-R improves to 0.83 ($+15.3\%$ over baseline)—by directly optimizing summary-level objectives that token-level cross-entropy cannot enforce. GRPO synergizes with weighted SFT rather than acting as a standalone boost. Combining all components yields the best trade-off: ITA-R $0.72 \rightarrow 0.87$ ($+20.8\%$) and GPT-4o overall $3.86 \rightarrow 4.13$ ($+7.0\%$), exceeding any individual component (Table 3).

6 Discussion

Token-level weighting and sequence-level rewards jointly target the two core failure modes. In Stage 1, keyword-aware and visual-alignment weighting counter key-information omission and cross-modal hallucination by raising the gradient on salient entities and visually grounded spans. Stage 2 reinforces the same two axes at the sequence level: the keyword-coverage reward penalizes missing entities, and the image-text alignment reward suppresses ungrounded visual mentions, especially on image-dominant documents.

Composite reward balances coverage and faithfulness without sacrificing conciseness or coherence. Optimizing a single reward in isolation over-shoots one axis at the cost of others: keyword coverage alone inflates length with peripheral entities, and image-text alignment alone encourages indiscriminate visual mentions. Coupling these signals with ROUGE and a length penalty lets the four components mutually regularize, yielding summaries that are informative, visually faithful, concise, and coherent.

7 Conclusion

We study multimodal long-document summarization under long-context and cross-modal evidence sparsity, focusing on key-information omission and cross-modal hallucination. We introduce MMLDSum-Bench, a multi-domain, multi-length, multi-ratio benchmark, and propose MMLDSum-LLM, a two-stage recipe that combines weighted SFT (visual alignment and keyword awareness) with GRPO using verifiable, multi-objective rewards. Across automatic and judge-based evaluations, MMLDSum-LLM improves key-information coverage and cross-modal consistency compared with SFT-only baselines. Future work includes stronger chart-specific grounding, adaptive reward re-weighting conditioned on length/ratio, and more reliable multimodal evaluation protocols.

Limitations

First, our rewards still rely on proxy signals (e.g., ROUGE, keyword coverage, and image-text alignment) that can miss fine-grained factual errors or chart-specific reasoning, especially for dense plots and complex diagrams. Second, long-context behavior remains fragile: when key evidence is sparse and distributed across distant sections, the model may still omit crucial details or overfit local evidence despite weighted training. Third, evaluation costs remain high because judge-based scoring and atomic-claim verification are computationally expensive, which limits large-scale ablations and rapid iteration. Finally, our benchmark focuses on static documents with pre-extracted images; extending to dynamic or interactive visuals (e.g., videos or embedded charts with underlying data) remains future work.

Reproducibility

To support full reproducibility and community adoption, we will publicly release: (i) **MMLDSum-Bench**, including all benchmark documents, paired reference summaries, and split metadata (SFT/R-L/test); (ii) **training code** for both Stage 1 (visual-alignment and keyword-aware weighted SFT) and Stage 2 (GRPO with multi-objective verifiable rewards), together with training configuration files and hyperparameter settings used in all reported experiments; (iii) **evaluation code**, covering the full automated evaluation suite—LLM-as-a-judge prompts (GPT-4o and GPT-5 five-dimension scoring), atomic-claim extraction and verification pipelines, ITA-R computation (BGE-M3 with threshold 0.65), and ROUGE scoring; (iv) **model checkpoints** for all reported MMLDSum-LLM variants (3B, 7B, 8B); and (v) all **inference prompt templates** used during model evaluation. All training runs use fixed random seeds. We will document software versions (Python, PyTorch, Transformers, vLLM) and hardware specifications (NVIDIA H20 $\times$ 8). Benchmark data is filtered to remove personally identifiable information, and all source dataset licenses are respected.

Use of AI Assistants

The AI assistant, GPT-4o, is used solely for refining the writing of our paper.

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2309.07597*.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and 1 others. 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.

Akash Ghosh, Mohit Tomar, Abhisek Tiwari, Sriparna Saha, Jatin Salve, and Setu Sinha. 2024. From sights to insights: Towards summarization of multimodal clinical documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13117–13129.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14867–14878.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2025. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3599–3607.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Computing Surveys*, 55(13s):1–36.

Abdullah Faiz Ur Rahman Khilji, Utkarsh Sinha, Pintu Singh, Adnan Ali, Sahinur Rahman Laskar, Pankaj Dadure, Riyanka Manna, Partha Pakray, Benoit Favre, and Sivaji Bandyopadhyay. 2023. Multimodal text summarization with evaluation approaches. *Sādhanā*, 48(4):226.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM computing surveys*, 55(8):1–35.

Atharva Kumbhar, Harsh Kulkarni, Atmaja Mali, Sheetal Sonawane, and Prathamesh Mulay. 2023. The current landscape of multimodal summarization. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 797–806.

Oliver Langston and Brian Ashford. 2024. Automated summarization of multiple document abstracts and contents using large language models. *Authorea Preprints*.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, and 1 others. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158.

Yinhong Liu, Jianfeng He, Hang Su, Ruixue Lian, Yi Nian, Jake Vincent, Srikanth Vishnubhotla, Robinson Piramuthu, and Saab Mansour. 2025. Mdseval: A meta-evaluation benchmark for multimodal dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14707–14727.

Ming Lu, Yang Liu, and Xiaoming Zhang. 2024. A modality-enhanced multi-channel attention network for multi-modal dialogue summarization. *Applied Sciences*, 14(20):9184.

Qiduo Lu, Chenhao Zhu, and Xia Ye. 2022. Research on multimodal summarization by integrating visual and text modal information. In *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pages 882–889. IEEE.

Louis Mahon and Mirella Lapata. 2024. A modular approach for multimodal summarization of tv shows. *arXiv preprint arXiv:2403.03823*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, and 1 others. 2024. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21921.

Shaik Rafi and Ranjita Das. 2024. Sct: summary caption technique for retrieving relevant images in alignment with multimodal abstractive summary. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–22.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*.

Zusheng Tan, Xinyi Zhong, Jing-Yu Ji, Wei Jiang, and Billy Chiu. 2025. Enhancing large language models for scientific multimodal summarization with multimodal output. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 263–275.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning. *arXiv preprint arXiv:2505.17667*.

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

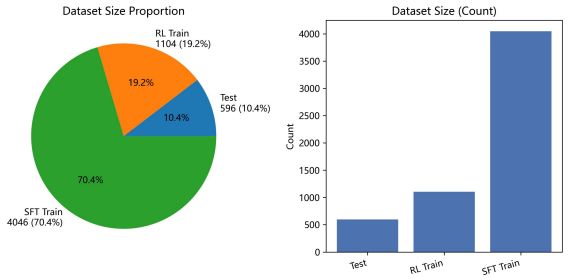
Zekun Yang, Jiajun He, and Tomoki Toda. 2024. Multimodal video summarization based on two-stage fusion of audio, visual, and recognized text information. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6. IEEE.

Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2025. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *Preprint*, arXiv:2403.02901.

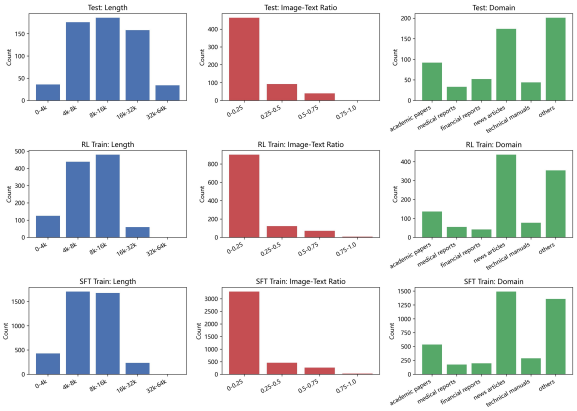
A Dataset Construction and Statistics

A.1 MMLDSum-Bench Statistics (Illustrative Figure)

Figure 4 summarizes the data composition and distribution patterns of MMLDSum-Bench. The split



(a) Dataset size and proportion.



(b) Length, ratio, and domain distributions.

Figure 4: MMLDSum-Bench statistics for the SFT, RL, and test splits.

size overview highlights the relative scale of the SFT, RL, and test sets. The distribution grid reports (i) length bins following the paper's standard 4k/8k/16k/32k/64k ranges, (ii) image–text ratio buckets at 0.25/0.5/0.75/1.0, and (iii) the six-domain taxonomy used in the paper: academic papers, medical reports, financial reports, news articles, technical manuals, and others. Together, these statistics validate that the benchmark spans diverse domains and modality balances while remaining focused on long-context settings.

A.2 Pipeline for Automatic Summary Construction

A.3 Data Quality Validation Protocol and Results

A.3.1 Annotator Profiles

Three annotators (A1, A2, A3) participate in the quality validation. All hold graduate-level degrees in natural language processing or related fields and have at least two years of research experience with text summarization and multimodal document understanding. Prior to formal annotation, all annotators complete a calibration session on 30 pilot samples (excluded from the final evaluation set) to align scoring criteria and resolve ambiguities in dimension definitions.

A.3.2 Stratified Sampling Protocol

We draw 600 summaries from the benchmark via stratified sampling along three axes to ensure representative coverage:

- **Domain**: samples are allocated proportionally across six domains (academic papers, medical reports, financial reports, news articles, technical manuals, and others).
- **Context-length scale**: samples are drawn from all five length bins (4K, 8K, 16K, 32K, 64K tokens) with proportional allocation reflecting the benchmark distribution, while enforcing a minimum of 30 samples per bin.
- **Visual-textual modality distribution**: samples cover all four modality categories (heavily text-dominant, lightly text-dominant, lightly image-dominant, heavily image-dominant), with oversampling applied to minority categories to ensure at least 20 samples per category.

A.3.3 Annotation Scheme

Each of the 600 summaries is independently scored by all three annotators on five dimensions using a 1–5 Likert scale:

- **Completeness**: whether the summary covers all salient information from the source document across modalities.
- **Accuracy**: whether the factual claims in the summary are correct and free of hallucinations.
- **Coherence**: whether the summary is logically organized and easy to follow.
- **Conciseness**: whether the summary avoids redundancy and unnecessary detail.
- **Overall quality**: a holistic assessment of the summary.

In addition, 200 atomic claims are randomly sampled from the generated summaries. Each claim is independently verified by all three annotators against the source document (text and associated images) and labeled as *Supported*, *Partially Supported*, or *Unsupported*.

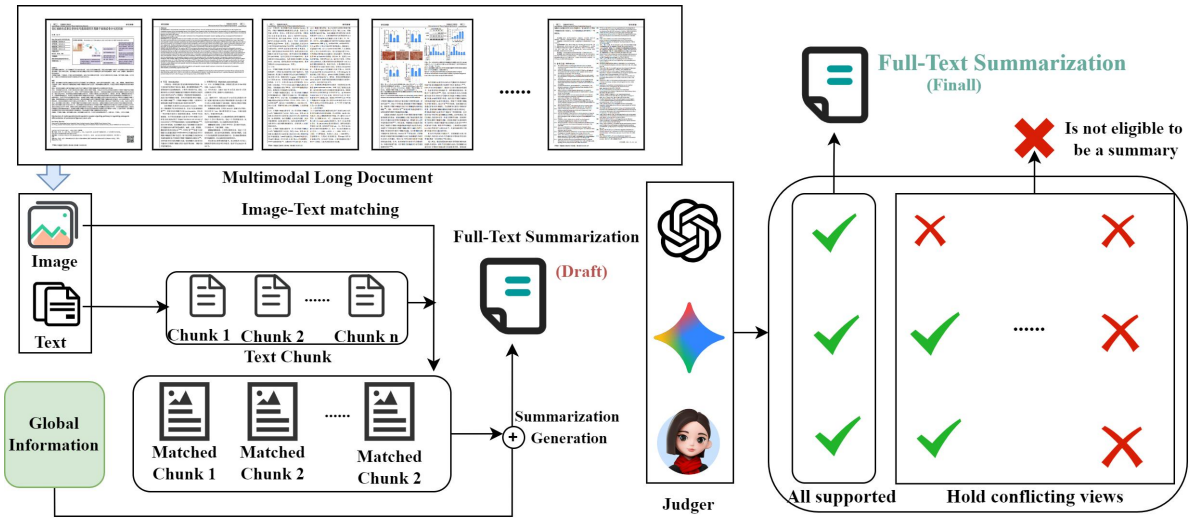


Figure 5: Pipeline for MMLDSum-Bench construction: image–text matching and chunking, global information fusion to draft summaries, and multi-model judging to select the final summary.

Table 4: Per-dimension human evaluation results on 600 stratified summaries.

Dimension	Mean	95% CI	Cohen's κ
Completeness	4.58	[4.52, 4.64]	0.80
Accuracy	4.82	[4.78, 4.86]	0.86
Coherence	4.75	[4.70, 4.80]	0.84
Conciseness	4.68	[4.62, 4.74]	0.82
Overall	4.70	[4.65, 4.75]	0.83

Table 5: Claim-level manual verification results on 200 atomic claims.

Metric	Value
Supported rate	88.5%
Partially supported rate	7.0%
Unsupported rate	4.5%
Annotator agreement (Fleiss' κ)	0.81

A.3.4 Detailed Annotation Results

Per-dimension summary-level evaluation. Table 4 reports per-dimension mean scores with 95% bootstrap confidence intervals and pairwise Cohen's Kappa averaged over the three annotator pairs.

Claim-level manual verification. Table 5 reports the results of claim-level verification on 200 atomic claims, along with inter-annotator agreement measured by Fleiss' Kappa.

Regeneration-effect analysis. Table 6 quantifies the effect of the regeneration module by comparing quality scores before and after regeneration for the subset of samples that triggered the quality-threshold filter.

Table 6: Effect of the regeneration module on filtered samples.

Dimension	Before	After	Δ
Completeness	3.42	4.51	+1.09
Accuracy	3.78	4.76	+0.98
Coherence	3.85	4.70	+0.85
Conciseness	3.90	4.62	+0.72
Overall	3.56	4.65	+1.09
Pass rate	37.2% $\rightarrow$ 91.8%		

The results confirm that the dataset maintains high annotation quality across all dimensions (overall mean: 4.7/5.0; overall Cohen's κ: 0.83). Claim-level verification indicates a low unsupported-claim rate (4.5%), and the regeneration mechanism yields substantial quality improvements (average score increase of +0.95 across dimensions; pass-rate improvement from 37.2% to 91.8%).

A.4 Prompts for Automatic Summary Construction

The following prompt templates are used in the three-stage automatic summary construction pipeline described in Section 3.

Chunk-Level Summary Generation Prompt (System)

```
Role:
```

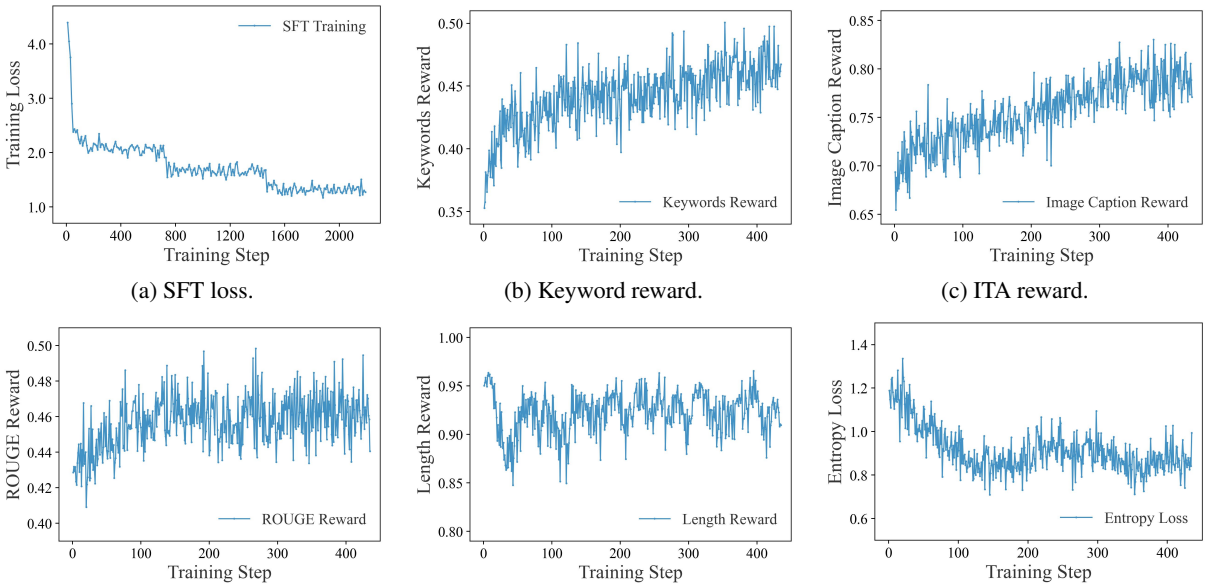
```
You are an expert in full-information
multimodal summarization. Generate summaries
in Simplified Chinese. Your core objective is
to preserve all source information, align
correctly with image positions, jointly
present text and image content, and keep the
summary logic/order exactly consistent with
the source.

Task background and objective:
- Input data: The user provides source text
and an image list. In the source text, <image
X> (X is a number) denotes an image marker.
Marker order matches the image list order
one-to-one (e.g., <image 1> corresponds to
image 1 in the list).
- Summary requirement: You must summarize both
source text and all images. Do not omit any
textual details (background, causes, process,
conclusions, opinions, definitions, features,
data, time, cases, etc.) or image information.
Do not repeat content.
- Image recall requirement: The summary must
recall all source images. Keep all markers
like <image 1> at their corresponding source
positions. Do not modify or delete these
markers. Ensure every marker in the summary
has a matching image in the source.

Summary rules:
1. Completeness:
- Reproduce details sentence by sentence so
users can recover all source information
without loss.
- Fully extract image information by image
type (chart/diagram/scene/flowchart/text-in-
image), including key elements, data,
relations, and scene descriptions; integrate
naturally at corresponding positions.
2. Accuracy:
- All content (events, opinions, data, time,
wording) must come from the source. No
fabrication.
- Keep critical wording exactly consistent
with the source (e.g., if the source says
"less than 3%", do not rewrite it as "only 3%"
or "more than 3%").
- Preserve relative time expressions (e.g.,
"this year", "last month", "the first half of
the year"); do not convert them into absolute
dates.
- Keep summary logic and order exactly
consistent with the source.
3. Image content presentation:
- Text and image information should have equal
importance, both presented completely in
source order with natural transitions.

Notes:
1. Do not delete or modify any <image X>
markers.
2. Do not reorder source content or logic.
3. Do not fabricate non-source content (text
details, image info, data, or opinions).
4. Do not output non-summary notes (e.g.,
"Image details are integrated above.").
5. Do not simplify key source details.
```

Global Information Extraction Prompt (System)

```
Role:
You are an information extraction specialist.
Extract document-level global information from
the given document and output it in the
required JSON format to support downstream
summarization.

JSON fields (must include all):
- "topic": one-sentence summary of the
document's core topic
- "outline": a list of major section/paragraph
titles
- "key_entities": repeatedly appearing key
entities, including but not limited to people,
organizations, locations, products,
technologies, and concepts

Example output (strict JSON, no extra
characters):
{
  "topic": "Global AI chip market analysis for
  Q3 2024",
  "outline": ["Market overview", "Major vendor
  updates", "Technology trends", "Outlook"],
  "key_entities": ["NVIDIA", "AMD", "H100",
  "compute power"]
}
```

Global Summary Generation Prompt (System)

```
Role:
You are a full-information replication
summarization expert. You can process both
text and image content jointly and generate
summaries in Simplified Chinese. The summary
must be complete and accurate, with logic and
order exactly consistent with the source.

Task background and requirements:
- Input data: The user provides global
document information and multiple chunk
summaries.
  a) Global information includes topic,
  outline, and key entities, which helps
  reconstruct the source structure and avoid
  fragmented writing.
  b) In each chunk summary, <image X> marks
  images, and all markers have been globally
  reindexed in document order.
- Summary requirement: Summarize all chunk
summaries and global information together,
preserving content and order exactly, with no
omission and no repetition.

Summary rules:
1. Structure: Use introduction - detailed
bullet points - optional conclusion.
2. Content:
   a) Completeness: Reproduce all details from
   chunk summaries (background, causes,
   process, conclusions, definitions,
   features, data, time, cases, etc.).
   b) Accuracy:
   - No fabricated content.
```



(a) SFT loss. (b) Keyword reward. (c) ITA reward.

(d) ROUGE reward. (e) Length reward. (f) Entropy (exploration).

Figure 6: Training curves for SFT loss and GRPO reward signals.

```
   - Keep key wording consistent with source
   values and semantics.
   - Preserve relative time expressions.
   - Keep logic/order exactly consistent with
   the source.
   c) Image content:
   - Summarize both text and image content in
   source order.
   - Extract key image elements/data/relations
   and integrate naturally.
3. Format: Any information coming from images
must be wrapped with
<image_tag>...</image_tag>.

Workflow:
Step 1: Write one introductory paragraph
summarizing text and images.
Step 2: Expand in ordered bullet points
according to source sequence, using global
topic/outline/entities to improve coherence
between chunks. All image-derived content must
be wrapped by <image_tag>...</image_tag>.
Step 3: Self-check coverage, factual
consistency (especially data/time), and order
consistency.
Step 4: If needed, add a final concluding
paragraph.

Important constraints:
1. Do not reorder chunk summaries.
2. Keep text and image information balanced.
3. Ensure all content is source-grounded and
consistent.
4. Do not output explanatory meta text.
5. Do not output raw image markers such as
<image x> in the final summary.
```

B Training Curves

Figure 6 visualizes the optimization dynamics across supervised fine-tuning and GRPO. The SFT loss decreases steadily, while reward components (keyword, caption, ROUGE, and length) rise as the policy improves. The entropy curve indicates exploration during RL, which stabilizes as rewards converge.

C Prompts for Inference and Evaluation

This appendix presents the prompt templates used for model inference and automated evaluation.

Inference-Time Summary Generation Prompt (User)

```
You are an expert in multimodal long-document
summarization. Your task is to generate a
summary in Chinese for a multimodal long
document. The summary must be complete,
accurate, and follow the same logical order as
the source.

Task background:
The user provides source text and an image
list. In the source text, <image X> (X is a
number) is an image marker. Marker order maps
one-to-one to the image list (e.g., <image 1>
corresponds to image 1).
You must summarize both text and image
content, and keep their presentation order
exactly aligned with the source.

Summary requirements:
Use an introduction - detailed bullet points -
optional conclusion structure:
1. Opening: one paragraph that gives a
high-level overview of text and images;
```

```
2. Middle bullet points: expand in detail
according to source order and paragraph
hierarchy, accurately presenting each part's
key content;
3. Ending (optional): one paragraph
summarizing the main idea, core conclusions,
and overall information.

=== Source Document Start ===
{article}
=== Source Document End ===

Now generate the summary based on the document
and images. Output only the summary, and do
not output any irrelevant content.
```

Five-Dimension LLM-as-a-Judge Prompt (System)

```
Role:
You are a precise and professional image-text
summary evaluator specialized in scoring
Chinese summaries generated from text+image
inputs. You provide rigorous step-by-step
analysis and quantitative scores.

Task and output:
- Input includes source text, image list, and
generated summary. <image X> markers in source
map one-to-one to the image list.
- Score each dimension from 1 to 5:
completeness, accuracy, conciseness,
coherence, and overall quality.
- Output must include two parts:
  1) detailed reasoning process for each
  dimension;
  2) final JSON scores for automatic
  extraction.

Scoring dimensions:
1. Completeness: no missing core text info or
key image info.
2. Accuracy: no factual deviation,
contradiction, or fabrication in text/image
descriptions.
3. Conciseness: no irrelevant content,
redundancy, or repeated statements.
4. Coherence: clear ordering and logical flow
consistent with source text-image structure.
5. Overall quality: holistic quality
considering all dimensions.

Output format:
[Reasoning]
... detailed analysis for each dimension ...
[Scores] (JSON only for scores)
{"completeness": [score], "accuracy": [score],
"conciseness": [score], "coherence": [score],
"overall": [score]}

Important:
- In the final JSON, output numeric values
only (e.g., 1, 2, 3, 4, 5), without units or
extra text.
```

Five-Dimension LLM-as-a-Judge Prompt (User)

```
=== Source Document Start ===
{article}
=== Source Document End ===

=== Summary Start ===
{summary}
=== Summary End ===
```

Atomic-Claim Extraction Prompt (System)

```
Your task is to extract all independent atomic
factual claims from the provided Chinese
summary text. An atomic claim is the smallest
complete statement that can be judged true or
false.

Strict rules (must be followed 100%):
1. One sentence, one fact: each claim must
contain exactly one independent fact.
2. Explicit information only: do not add
inference, external knowledge, assumptions,
interpretation, or opinion.
3. Preserve details: keep all dates, numbers,
amounts, named entities, acronyms, and
specific descriptions unchanged.
4. Split compound statements connected by
words such as "and/or/also/includes" into
multiple independent claims.
5. Split modifier-bearing facts into
independent atomic claims when modifiers carry
standalone facts.
6. Format each claim as a complete declarative
sentence with proper punctuation.
7. No omission and no duplication.
8. Output must be a single valid JSON string
only, with no prefix/suffix text.
   - Key name must be exactly: atomic_claims
   - No extra keys
   - Array elements must be JSON strings
   - Use ASCII JSON punctuation only

Required output format:
{"atomic_claims": ["Atomic claim 1.", "Atomic
claim 2.", "Atomic claim 3."]}
```

Atomic-Claim Extraction Prompt (User)

```
=== Summary Start ===
{summary}
=== Summary End ===
```

Atomic-Claim Verification Prompt (System)

```
You are a factual verification expert.
Determine whether each atomic factual claim is
supported by the summary.

Decision rule:
- true: the summary explicitly contains or
directly supports the claim
- false: the summary does not mention the
claim or contradicts it

Output format (JSON only, no extra text):
```

```
{"results": {"1": true, "2": false, "3":
true}}

Notes:
1. Output only claim IDs and boolean
judgments; do not output claim text.
2. IDs must align one-to-one with the input
claim numbering.
3. You must return judgments for all input
claims.
```

Atomic-Claim Verification Prompt (User)

```
Summary:
{summary}

Atomic claims:
{claims}

Output the support judgment for each
atomic-claim ID.
```

D Additional Statistics

This section presents a detailed analysis of model performance stratified by context-length bin across five metric families. The heatmaps in Figures 10–11 visualize per-model, per-length-bin performance, complementing the aggregate scores in Table 2 and providing finer-grained insight into how summarization quality degrades (or is maintained) under increasing document length. Length bins follow the standard 4k/8k/16k/32k/64k token ranges, and each cell reports the average score for all test documents in that bin.

Atomic-claim F1 (Figure 7). Atomic-claim F1 is sensitive to document length: most models exhibit a clear downward trend as context length grows from 4k to 64k tokens, although the decrease is not strictly monotonic for every system. The drop is most pronounced for weaker open-source baselines without explicit key-information training (e.g., Phi-4-Multimodal: $0.30 \rightarrow 0.16$; LongWriter-GLM4-9B: $0.62 \rightarrow 0.36$), and models with a 32k context ceiling (Step-1o-Vision-32K, InternVL3.5-14B/38B) likewise degrade visibly beyond 16k tokens; Step-1o-Vision-32K does not produce a result in the 64k bin due to forced truncation. MMLDSum-qwen3vl-8b is the strongest open-source system in every length bin $(0.87, 0.82, 0.78, 0.79, 0.79)$, surpassing both its backbone-matched SFT baseline (Qwen3-VL-8B-sft: $0.77, 0.75, 0.71, 0.74, 0.76$) and the CoD variant (Qwen3-VL-8B + CoD: $0.72, 0.72, 0.68, 0.68, 0.69$) across all bins, and remains stable around 0.78–0.79 in the 16k–64k range, suggesting that keyword-aware weighted

Figure 7: Atomic-claim F1 on MMLDSum-Bench. Higher values indicate better performance.

SFT together with GRPO's sequence-level keyword-coverage reward helps mitigate the difficulty of evidence selection in longer documents. A small number of systems instead exhibit non-monotonic or mildly increasing F1 with length (e.g., Qwen3-VL-32B: 0.58 at 4k vs. 0.70 at 64k; Qwen2.5-VL-72B: $0.44 \rightarrow 0.52$), indicating that document length alone is not the sole determinant of atomic-claim quality and that each model's specific long-context behavior also plays a role.

ITA-R (Figure 8). ITA-R exhibits the strongest sensitivity to the visual-textual modality distribution and document length of any metric in our suite. For most models, ITA-R is notably higher in the 4k–8k bin (where visual evidence is densely concentrated and spatially close to its textual descriptions) than in the 32k–64k bin (where images are scattered across distant document sections). This degradation is particularly sharp for models without explicit visual-alignment training, confirming the theoretical motivation of our visual-alignment weighted loss. MMLDSum-qwen3vl-8b consistently achieves the highest ITA-R across all length bins, and uniquely *improves* from the 8k to 16k bin for most document types — a pattern not observed in any baseline — suggesting that the GRPO image-

Figure 8: ITA-R on MMLDSum-Bench. Higher values indicate better performance.

Figure 9: ROUGE-L on MMLDSum-Bench. Higher values indicate better performance.

text alignment reward is especially effective when there is sufficient context for the model to identify image–text correspondences.

ROUGE-L (Figure 9). ROUGE-L generally decreases with document length: most closed-source systems peak in the 4k–8k bins and drop toward 64k (e.g., Qwen-VL-Max $0.37 \rightarrow 0.23$, Claude-4-Sonnet $0.35 \rightarrow 0.27$, GPT-5 $0.37 \rightarrow 0.29$), reflecting the difficulty of preserving lexical overlap when salient evidence becomes sparser. A few models are notably flatter (Doubao-Seed-1.6: 0.40 at 4k–16k, 0.34 at 64k; Qwen3-VL-Plus: $0.33 \rightarrow 0.30$). MMLDSum-qwen3vl-8b achieves the highest ROUGE-L in every length bin among open-source models, with a U-shaped profile $(0.39, 0.39, 0.36, 0.36, 0.40)$ that is robust at both ends. The advantage is most pronounced in the 64k bin, where it (0.40) surpasses Doubao-Seed-1.6 (0.34), Qwen3-VL-Plus (0.30), GPT-5 (0.29), and Claude-4-Sonnet (0.27). Compared with the SFT baseline (Qwen3-VL-8B-SFT: $0.25 \rightarrow 0.31$), our full two-stage model lifts ROUGE-L by 0.07–0.14 across all bins, indicating that the GRPO ROUGE reward and length penalty contribute substantial gains beyond weighted SFT alone in the most challenging long-context settings.

GPT-4o and GPT-5 judge scores (Figure 10 and Figure 11). Across both judges, performance degradation with length is pronounced for most open-source models but moderate for top closed-source models and our trained models. Closed-source models with 256k context windows (Qwen3-VL-Plus, Doubao-Seed-1.6, Qwen-VL-Max) maintain relatively stable GPT-4o scores across all five length bins, confirming that long-context ingestion capacity is a primary bottleneck for completeness. In contrast, models with 32k context limits (Step-1o-Vision-32k, InternVL3.5-14B/38B) exhibit a clear performance drop in the 32k–64k bin; Step-1o-Vision-32k in particular shows notably depressed completeness scores in the longest bin due to forced document truncation. GPT-5 scores generally follow the same trend as GPT-4o but with lower absolute values and wider inter-model gaps, particularly on the completeness and overall dimensions. MMLDSum-qwen3vl-8b achieves GPT-4o overall scores competitive with Claude-4-Sonnet and Step-1o-Vision-32k across the 4k–32k range, and maintains this level into the 32k–64k bin, demonstrating that the two-stage training framework successfully extends the effective summarization range of the 8B model.

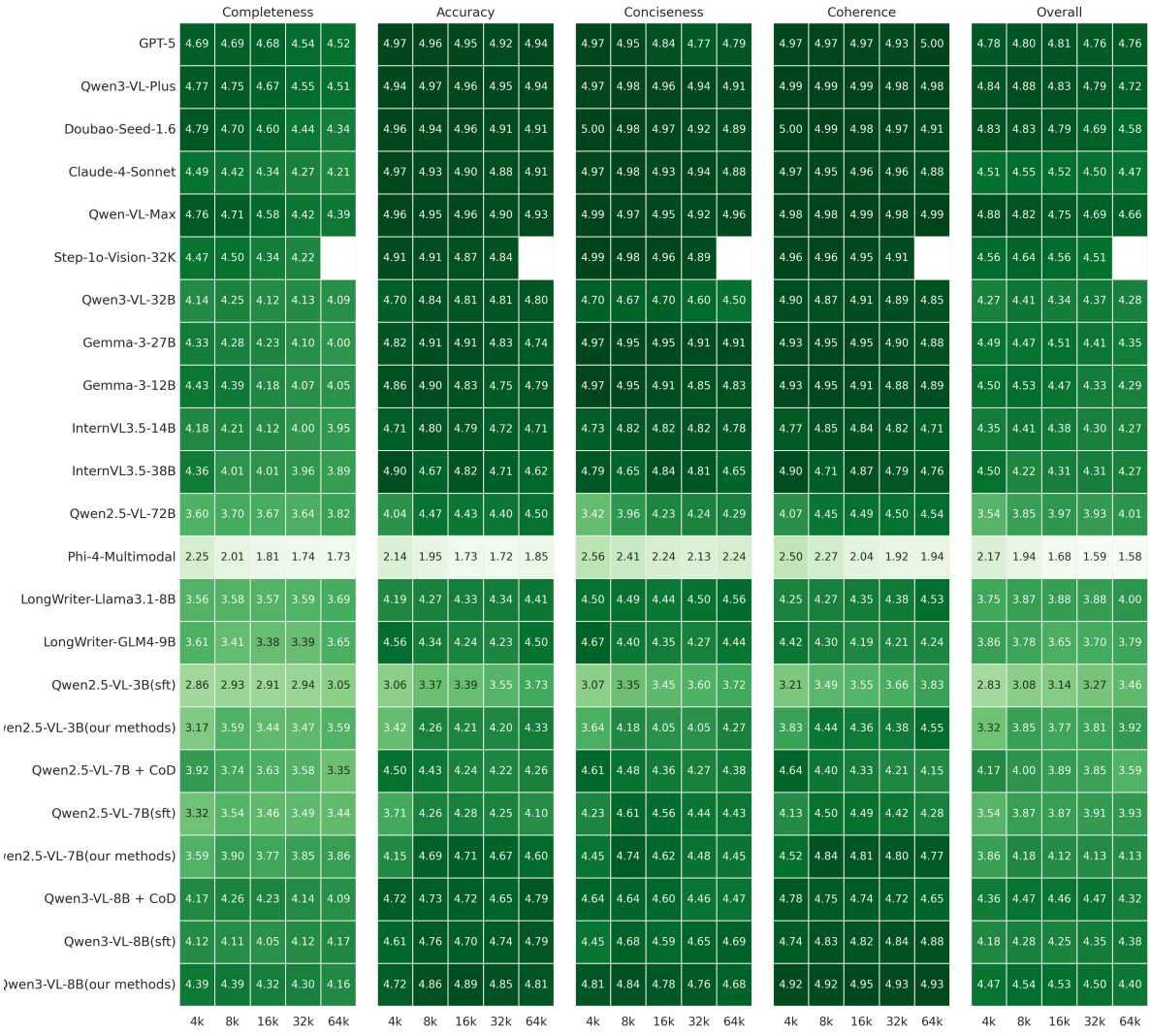


Figure 10: Heatmap visualization of GPT-4o judge-based scores on MMLDSum-Bench. Higher values indicate better performance.

Table 7: Cross-judge trend-consistency statistics between GPT-4o and GPT-5 on the 24 models in Table 2. Spearman ρ and Kendall τ-b measure rank agreement; Pearson r measures linear agreement. All reported correlations are significant at $p < 0.001$.

Dimension	Spearman ρ	Kendall τ-b	Pearson r
Completeness	0.910	0.750	0.890
Accuracy	0.704	0.571	0.681
Conciseness	0.759	0.594	0.771
Coherence	0.810	0.664	0.828
Overall	0.894	0.776	0.870
Pooled (N=120)	**0.787**	**0.614**	**0.776**

Cross-judge trend-consistency quantification. To turn the qualitative observation that GPT-4o and GPT-5 follow the same ranking into a quantitative claim, we compute three correlation coefficients between the two judges on the 24 evaluated models in Table 2 (the GPT-5 self-evaluation row is excluded). Spearman ρ and Kendall τ-b directly measure rank agreement, while Pearson r measures linear agreement; results per dimension and pooled across all five dimensions are reported in Table 7 and visualized in Figure 12. All correlations are highly significant ($p < 0.001$). Completeness ($\rho = 0.910$) and Overall ($\rho = 0.894$) exhibit the strongest agreement, indicating that the two judges essentially

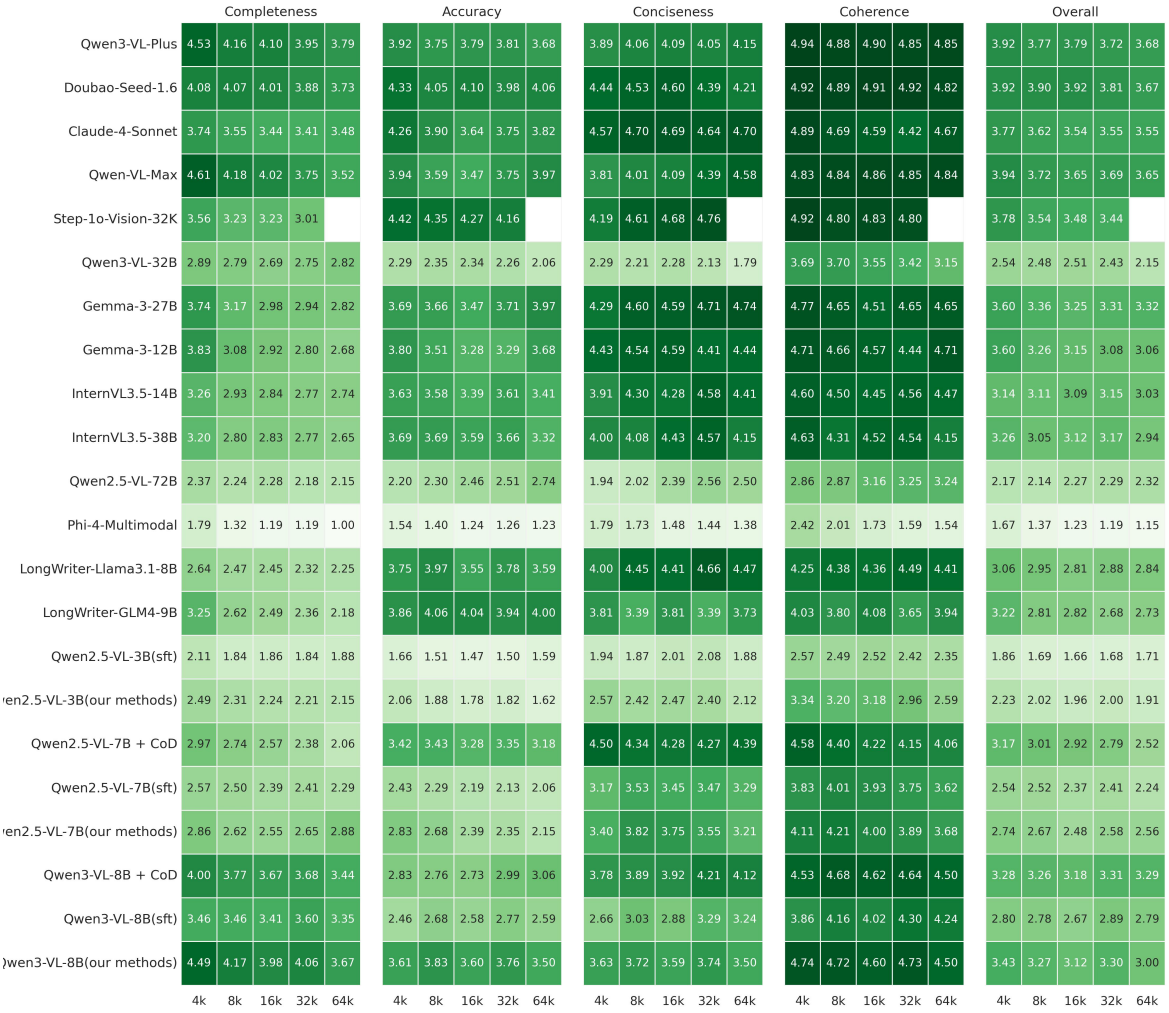


Figure 11: Heatmap visualization of GPT-5 judge-based scores on MMLDSum-Bench. Higher values indicate better performance.

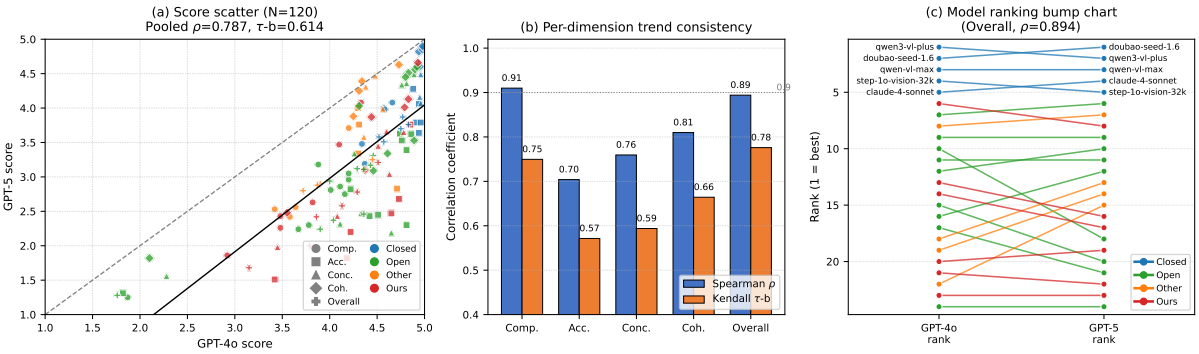


Figure 12: Quantitative cross-judge trend consistency between GPT-4o and GPT-5 on the 24 evaluated models in Table 2 (the GPT-5 self-evaluation row is excluded). **(a)** Score scatter pooled across all five dimensions; marker shape denotes the dimension, color denotes the model category, the dashed line is $y = x$, and the solid line is the least-squares fit. **(b)** Per-dimension Spearman ρ and Kendall τ-b; the dotted line marks the high-agreement threshold of 0.9. **(c)** Bump chart of model rankings: each line connects a model's GPT-4o rank to its GPT-5 rank on the Overall dimension; the top-ranked systems are nearly identical under both judges.

agree on which models are more complete and which are stronger overall; Accuracy is the least consistent ($\rho = 0.704$), consistent with GPT-5 applying a stricter standard on factual claims. The bump chart in Figure 12(c) further shows that the top-ranked systems are nearly identical under both judges, with only minor swaps within the top five. Together, these statistics confirm that the convergent trends reported above are not anecdotal: GPT-4o and GPT-5 produce trend-consistent rankings, supporting the validity of using both judges as complementary evaluation signals.

Cross-metric consistency and key takeaways. Aggregating the length-stratified results across the four metric families yields three consistent findings. First, *factual completeness (atomic-claim F1) and visual grounding (ITA-R) are the two metrics most sensitive to document length*, and both receive the largest absolute gains from MMLDSum-LLM's two-stage training relative to the backbone-matched SFT baseline and the CoD variant. Second, *the 32k–64k regime is the most discriminating*: systems constrained to a 32k context window (Step-1o-Vision-32K, InternVL3.5-14B/38B) degrade sharply or fail to produce outputs, whereas models with 128k+ context windows—including MMLDSum-qwen3vl-8b—retain competitive performance, confirming that adequate long-context ingestion capacity is a prerequisite for robust summarization on MMLDSum-Bench. Third, *cross-metric agreement supports the validity of the observed gains*: improvements on ITA-R and ROUGE-L (used as GRPO reward components) are accompanied by consistent improvements on the held-out evaluation signals—atomic-claim F1 and GPT-4o/GPT-5 judge scores—suggesting that the gains reflect genuine multidimensional quality improvement rather than reward fitting.