

LongWriter-V: Enabling Ultra-Long and High-Fidelity Generation in Vision-Language Models

Shangqing Tu*
DCST, Tsinghua University
Beijing 100084, China
tsq22@mails.tsinghua.edu.cn

Yucheng Wang*
DCST, Tsinghua University
Beijing 100084, China
wang-yc24@mails.tsinghua.edu.cn

Daniel Zhang-Li
DCST, Tsinghua University
Beijing 100084, China
zlnn23@mails.tsinghua.edu.cn

Yushi Bai
DCST, Tsinghua University
Beijing 100084, China
bys22@mails.tsinghua.edu.cn

Jifan Yu†
IoE, Tsinghua University
Beijing 100084, China
yujifan@mail.tsinghua.edu.cn

Yuhao Wu
Singapore University of Technology
and Design
Singapore, Singapore
wu_yuhao@mymail.sutd.edu.sg

Lei Hou†
DCST, Tsinghua University
Beijing 100084, China
houlei@tsinghua.edu.cn

Huiqin Liu
IoE, Tsinghua University
Beijing 100084, China
liuhq@tsinghua.edu.cn

Zhiyuan Liu
DCST, Tsinghua University
Beijing 100084, China
liuzy@tsinghua.edu.cn

Bin Xu†
DCST, Tsinghua University
Beijing 100084, China
xubin@tsinghua.edu.cn

Juanzi Li
DCST, Tsinghua University
Beijing 100084, China
lijuanzi@tsinghua.edu.cn

Abstract

Existing Large Vision-Language Models (LVLMs) can process inputs with context lengths up to 128k visual and text tokens, yet they struggle to generate coherent outputs beyond 1,000 words. We find that the primary limitation is the absence of long output examples during supervised fine-tuning (SFT). To tackle this issue, we introduce LongWriter-V-22k, a SFT dataset comprising 22,158 examples, each with multiple input images, an instruction, and corresponding outputs ranging from 0 to 10,000 words. Moreover, to achieve long outputs that maintain high-fidelity to the input images, we employ Direct Preference Optimization (DPO) to the SFT model. Given the high cost of collecting human feedback for lengthy outputs (e.g., 3,000 words), we propose IterDPO, which breaks long outputs into segments and uses iterative corrections to form preference pairs with the original outputs. Additionally, we develop MMLongBench-Write, a benchmark featuring six tasks to evaluate the long-generation capabilities of VLMs. Our 7B parameter model, trained with LongWriter-V-22k and IterDPO, achieves impressive performance on this benchmark, outperforming larger proprietary models like GPT-4o. Our models, data and code are available at: <https://github.com/THU-KEG/LongWriter-V>.

*Both authors contributed equally to this research.

†Corresponding authors.



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '25, Dublin, Ireland*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2035-2/2025/10

<https://doi.org/10.1145/3746027.3755221>

CCS Concepts

• Computing methodologies → Natural language generation.

Keywords

Cross-modal application, Multimodality

ACM Reference Format:

Shangqing Tu, Yucheng Wang, Daniel Zhang-Li, Yushi Bai, Jifan Yu, Yuhao Wu, Lei Hou, Huiqin Liu, Zhiyuan Liu, Bin Xu, and Juanzi Li. 2025. LongWriter-V: Enabling Ultra-Long and High-Fidelity Generation in Vision-Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3746027.3755221>

1 Introduction

Recent advancements in Large Vision-Language Models (LVLMs) have significantly enhanced their capabilities in processing visual and textual inputs [1, 48]. Notably, there have been substantial breakthroughs in the long-context capabilities of VLMs [30, 43]. For instance, Qwen2-VL [35] can now understand videos up to 20 minutes, with a context window of 32k tokens. This progress has significantly expanded the scope of tasks that VLMs can handle, making them more applicable to real-world scenarios.

However, despite the increased input context window, the effective output length of VLMs remains limited. To verify this limitation, we collect a benchmark comprising six tasks that require VLMs to generate long texts based on visual inputs (as shown in Figure 1). By adjusting the required output length in the instructions, we found that all existing models struggle to generate outputs exceeding 1,000 words (Section 2). In real-world scenarios, such long-output

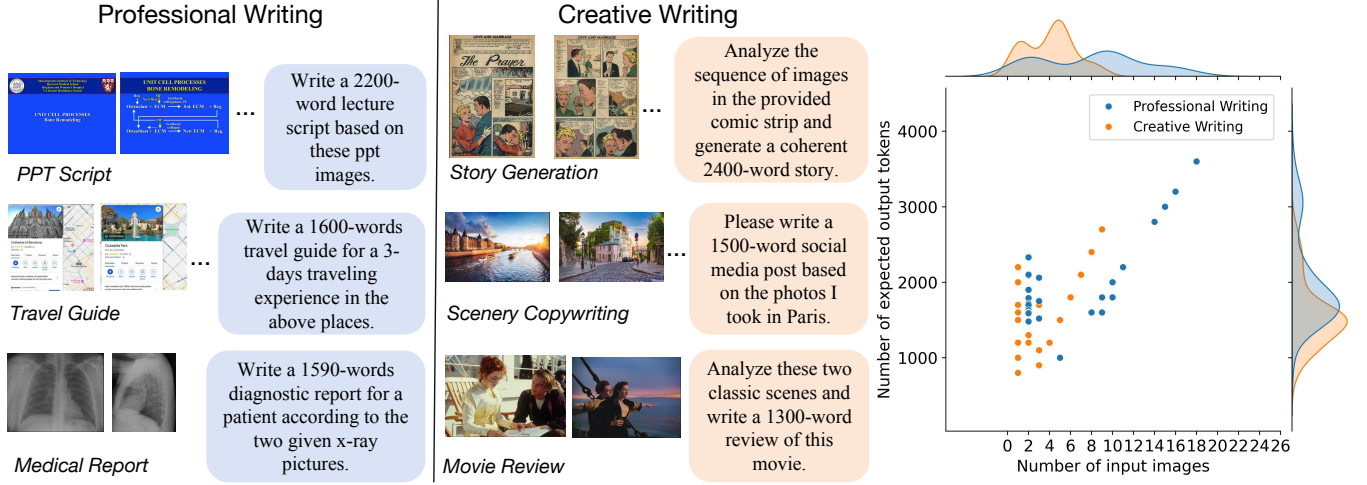


Figure 1: Left: Six examples for each type of task in MMLongBench-Write. They are divided into two categories: professional writing and creative writing. The former requires professional knowledge, while the latter does not. Right: The joint distribution of the number of input images and the expected output length for data in both categories. Most data requires a 1000+ word output with given images, challenging the long-generation capabilities of VLMs.

queries are common user demands [7]. For example, (1) creative writing tasks may require generating detailed stories or essays based on visual prompts [15], and (2) professional writing tasks may involve writing comprehensive reports or analyses from visual data [13]. To meet these practical needs, it is essential to enhance the long-output capabilities of VLMs.

To investigate the reasons behind the limited long-output capability of VLMs, we are inspired by the LongWriter [3], which adjusts the output length distribution of the supervised fine-tuning (SFT) data to observe changes in model output length. Our experiments revealed that the proportion of long-output examples in the SFT data determines the model’s output length. This finding explains why VLMs typically have an output length limit of around 1,000 words. Existing visual instruction tuning datasets [29], such as LLaVA [21], mainly contain tasks like grounding [22] and caption generation [34], with most outputs being less than 300 words [20].

To fill the gap, we select long-output instruction-image pairs from MMEval [24] as inputs. In addition to single-image inputs, we also constructed other forms of data, including multi-image inputs and backtranslated instructions [37], to enrich the diversity of the input data. To generate long outputs, we propose a plan-and-write approach: LongWriter-Agent-V. This method involves providing input images and writing instructions to GPT-4o to first generate an outline and then sequentially write the text in segments. Through this approach, we collect LongWrite-V-22k, a dataset of 22k long-output examples.

Using LongWrite-V-22k for SFT, the output length of Qwen2.5-VL-7B-Instruct [33] can be extended beyond 3,000 words. However, longer outputs often introduce issues such as repetition and hallucination [9]. To improve the fidelity of long outputs, we adapted the approach from RLHF-V [46], where human experts revise the model’s outputs to form preference pairs for Direct Preference Optimization (DPO). Since traditional DPO [28] is typically performed

on short texts of around 300 words, and LongWriter-V’s output length can exceed 3,000 words, the annotation cost is extremely high. To enhance the efficiency of preference data utilization, we proposed IterDPO, which divides long outputs into N segments, treating each segment’s revision as a preference pair. This method allows the model to learn fine-grained human corrections for each segment and effectively multiplies the use of a single long-output preference pair by N times. Through LongWriter-V-22k SFT and IterDPO, our 7B model achieves impressive performance in both output length and quality, surpassing powerful VLMs like GPT-4o.

In summary, our contributions are as follows:

- We construct MMLongBench-Write to evaluate the long-output capabilities of VLMs and find that the output length limit of existing VLMs is around 1,000 words.
- We collect the SFT dataset LongWrite-V-22k, enabling VLMs for 3,000+ word generation.
- We propose IterDPO, which effectively improves the text quality of long-output VLM.

2 Preliminaries

In the preliminary experiments, we first collect MMLongBench-Write, a benchmark with visual inputs and long-output requirements. Then, we conduct an evaluation on this benchmark to explore the maximum output length of VLMs. Besides, we reveal that the main reason for bounded output length lies in the length distribution of SFT data.

MMLongBench-Write. The ability to write long texts based on visual inputs is a fundamental skill in various real-world applications and can be broadly categorized into professional writing and creative writing, depending on whether specialized knowledge is required [31]. To evaluate how well that VLMs master the two skills, we design three specific tasks for each skill. For each task, we curate 20 representative instructions with input images as test data. To

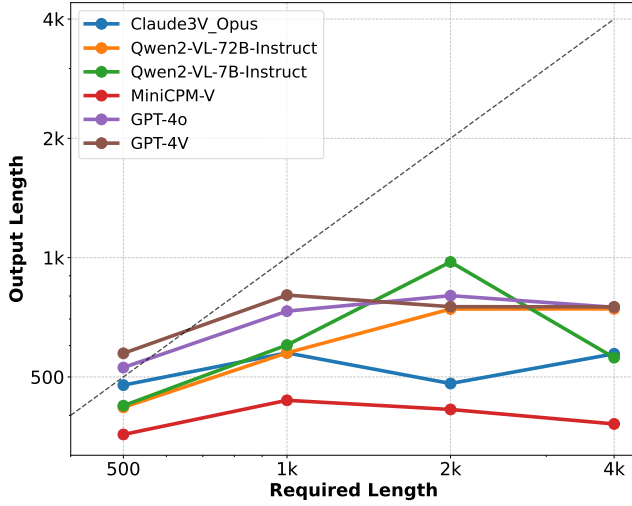


Figure 2: LongWriter-V-Ruler test across different output length requirements. The horizontal line show the overall upper bound for current VLMs.

ensure diversity, half instructions are in English and half are in Chinese. Figure 1 shows six examples of the benchmark and data distribution. It highlights that professional writing tasks typically involve more input images and require longer output lengths.

LongWrite-V-Ruler. To explore the maximum output length of VLMs, we select 8 examples from MMLongBench-Write benchmark, with four samples in English and four in Chinese. As depicted in Figure 1, each instruction is in the form of "Write an L -word article for the given pictures". We construct a diverse test set by changing the length requirement L . This test set uses $L \in \{500, 1000, 2000, 4000\}$, which consists of 32 test prompts in total.

Evaluation Result. We conduct the LongWrite-V-Ruler test on three open-source VLMs and three proprietary models. In Figure 2, we plot the required output length (x-axis) and the corresponding average output length (y-axis) for 12 instructions. We can observe that there exists an upper bound of 1000 output length for all models, which reveals their common weakness in writing long articles.

Preliminary Experiment. As the controlled experiments in LongWriter [3] has revealed that the maximum output length of LLM is correlated with the maximum output length of SFT data, we further explore how the average output length of SFT data can influence the long-generation capabilities of VLM. We fine-tune Qwen2-VL-7B-Instruct [35] on three visual instruction datasets sampled from our final SFT data. Each dataset has 10k examples with different average output length respectively (0.8k, 1.8k and 2.8k). Figure 3 shows the trained models' performance on LongWrite-V-Ruler, we observe that the model's maximum output length increases with the average output length of SFT data. Besides, we find that **the number of long-output examples** is crucial for extending the output length of VLMs. For example, the training set with an average length of 1.8k contains 1% data exceeding 4k output length, but the model trained on it fails to generate 4k tokens (orange line). In contrast, the model trained with 21% data exceeding 4k output length is able to do that (blue line). This result indicates that the

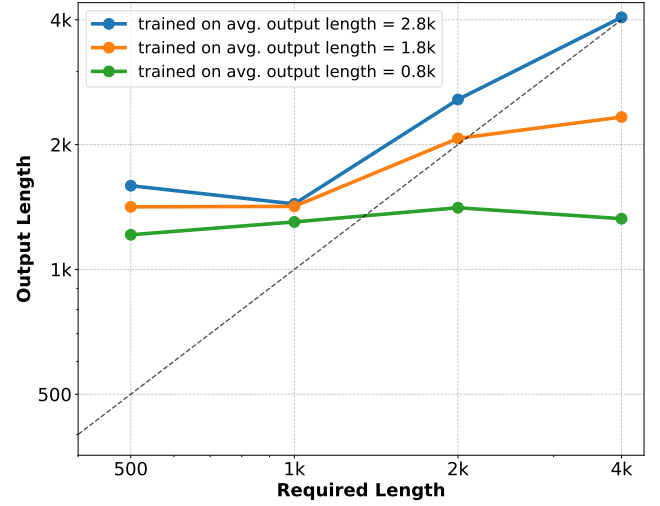


Figure 3: LongWriter-V-Ruler test for Qwen2-VL-7B-Instruct trained on 10k SFT data samples with different average output lengths.

main reason that limits the VLM's output length is lack of enough long-output examples in SFT data.

3 LongWriter-V: Data and Training

In this section, we will introduce the data collection and training process for unlocking the long generation capability of VLMs.

3.1 Data Collection

Figure 4 depicts the overall pipeline of our data collection process, which consists of two phases: SFT and DPO data collection.

3.1.1 SFT Data Collection. Existing VLMs fail to directly generate texts exceeding 1k tokens, so we develop a two-stage method to generate long texts as SFT data.

LongWrite Agent-V. Before introducing our method, we first formalize the task's objective. Given several input images v and an user instruction x , our goal is to generate a text y^* that aligned with user's length and quality requirements:

$$y^* = \arg \max_y (s_l(y) + s_q(y)) P_\theta(y|v, x) \quad (1)$$

where s_l and s_q is the scoring function for judging the length and quality of the output, respectively. P_θ is the function representing the end-to-end solution, while existing VLMs may not be directly applied as their maximum output lengths are below 1k. To utilize off-the-shelf VLMs, we propose a two-stage method for generating long texts. Inspired by the plan-and-write method from LongWriter [3], we first prompt the VLM to generate an outline o that structures the output, plans the content, and specifies the word count for each paragraph. This outline breaks down the complex long-output task into manageable sub-tasks. Next, we use the VLM to fill in each paragraph and concatenate them to form the final output:

$$y^* = \arg \max_o P_1(o|v, x) \arg \max_y (s_l(y) + s_q(y)) P_2 \quad (2)$$

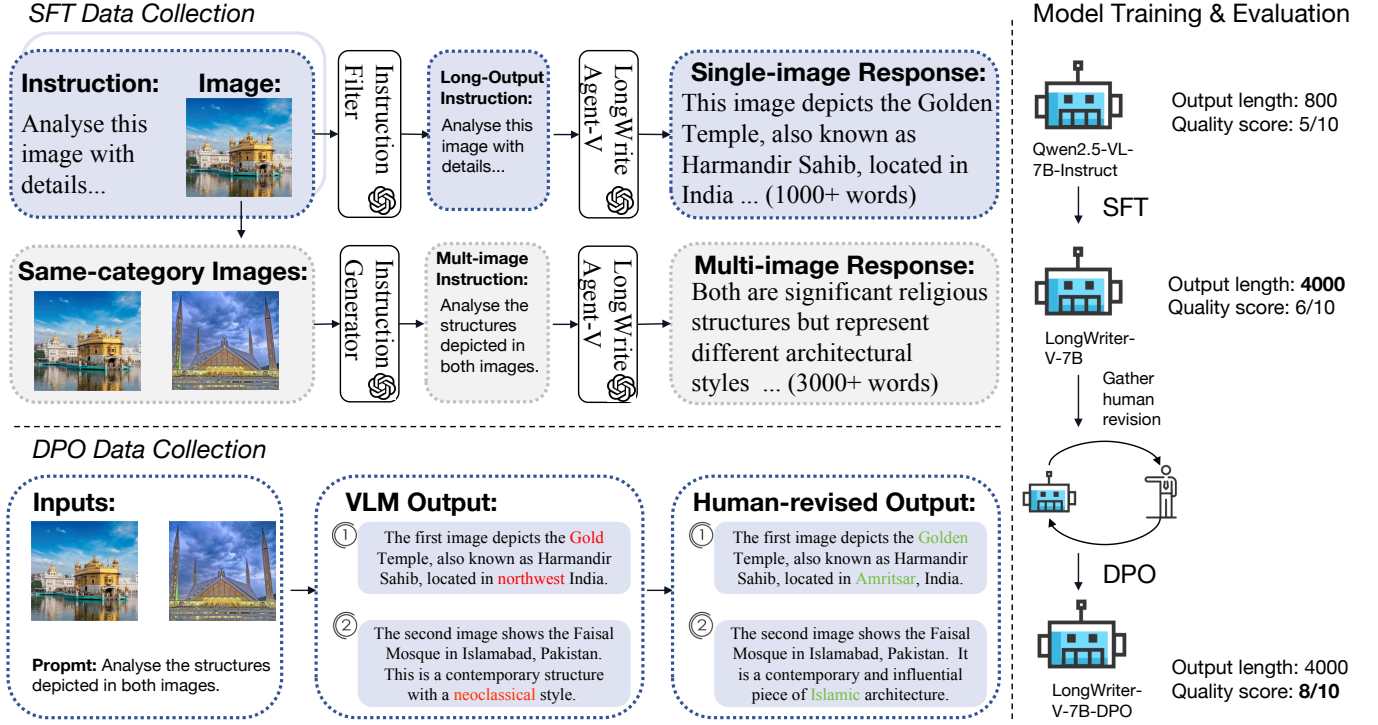


Figure 4: Data collection and model training pipeline of LongWriter-V. The SFT data includes both single-image and multi-image input for long text output. The DPO data contains human revision over each paragraph of VLM’s long output. We first SFT the VLM to extend its output length, then conduct iterative direct preference optimization to learn fine-grained human feedbacks.

$$P_2(y|v, x, o) = \prod_{i=0}^n p(y_i|v, x, o_i, y_{<i}) \quad (3)$$

where P_1 is the modeling function for first stage, which takes input images and instruction to write an n -paragraph outline $o = \{o_i, i = 1, \dots, n\}$. P_2 refers to the second stage, where the VLM outputs the content y_i paragraph by paragraph based on the input information, outline o_i and previous paragraphs $y_{<i}$. In practice, we design two detailed prompts for guiding VLM to implement the two stages.

Visual Instruction Collection. To collect long-output visual instructions for SFT, we choose MMEval [24] as our primary data source. MMEval is a large-scale, open-domain dataset containing 480k image-text instruction pairs, sourced from diverse datasets such as LLaVA-Instruct [21] and ShareGPT4V [4]. However, the average output length in this dataset is relatively short (54.85 tokens), necessitating a filtering process to identify long-output instructions. We first check the original response length of each example and select those with output length over 128, yielding 55,835 valid data. Next, we utilize GPT-4o to verify whether each instruction genuinely requires a long output and whether the associated image is sufficiently relevant to the instruction by using this prompt:

You will receive an image and an instruction from a user to an AI assistant, please determine whether the instruction requires the AI assistant to write an article for the given image, and the length of the article is more than 1,000 words in English (or 1,000 characters in Chinese). If the instruction does not

mention the word requirement, please determine whether the user’s intention of the response length is more than 1,000 words. If the instruction is unrelated with the image, please reply “no”. Instruction: {User Instruction}

Finally, we get 8,115 single-image instructions after filtering. **Multi-image Instruction Generation.** As the original data in MMEval only has one image for each instruction, we synthesize some multi-image instructions to increase the diversity of SFT data. We select three subsets of MMEval: wikiart, web-landmark, web-celebrity. Each subset contains hundreds of images in the same category. For example, images in web-landmark are all landmark pictures taken from different world attractions. We randomly sample 2 or 4 same-category images and then ask gpt-4o to generate an instruction that require long output for these images:

You will receive {Image Number} images and an instruction from a user to an AI assistant, this original instruction is targeted for the first image solely. Now please rewrite this instruction to a challenging long-output one that need using visual information from all the input images, and the length of the expected output should be more than 2,000 words in English (or 2,000 characters in Chinese). Here are three examples of challenging long-output instructions:
Example instruction 1: {Example Instruction 1}
Example instruction 2: {Example Instruction 2}

Example instruction 3: {Example Instruction 3}

Now, you should rewrite the following instruction:

Instruction: {User Instruction}

Please rewrite this user instruction to a challenging long-output instruction that requires the use of all the input images. Please output only the rewritten instruction, do not output other content.

We obtain 6,313 multi-image instructions in this way. Apart from synthetic data, we also collect natural multi-image data from an open-source PPT dataset, Zenodo10K [49]. We transform these slides into images as visual inputs and set the instruction as "Write a lecture script for these slides". We choose those slides that has at least 2 pages and at most 30 pages, resulting in 7,730 data.

Backtranslation. Through above processes, we collect 22,158 single-image and multi-image instructions in total. Using the LongWrite Agent-V pipeline, we generate long output for each visual instruction as SFT data. We call this training data LongWrite-V-22k. But most instructions don't specify the exact word count requirement, models trained on these data may lack the ability to follow the writing instruction with word count requirements. Therefore, we sample 5,000 data from LongWrite-V-22k and calculate the length of the output L then add a requirement "Please write L-word in total." to the end of the instruction and use gpt-4o-mini for rephrasing the instruction to maintain consistency. This is inspired by previous backtranslation [19] method on training long-output LLMs [27].

3.2 DPO

3.2.1 DPO Data Collection. The SFT data aims to extend VLMs' output length. But the longer outputs may bring more hallucinations and repetitions. So the follow up question is: *how to improve the generation fidelity of long output VLM?* Previous works often adapt direct preference optimization [23, 28] to correct the hallucinations of VLMs. We follow the data format in RLHF-V [46] which uses the segment-level corrections on VLM's outputs as feedback.

VLM Output Collection. To collect long responses, we select 100 slides that were not included in LongWrite-V-22k for VLM to generate scripts. These slides were previously used for teaching on MOOC platforms [45] and cover 10 subjects such as Computer Science, Math and Physics. Each subject may contain 4 to 16 slides and each slide may have 10 to 30 pages. We use LongWriter-V-7B, the VLM trained on our SFT data, to generate scripts for each slide. The long scripts are segmented by sections and aligned with each page of the given slide. We find that LongWriter-V-7B tends to output fewer sections than the number of total pages, which is one of the issues that we would ask human annotators to fix.

Human Revision Collection. To get high-quality feedback on the flawed output of SFT model, we hire 10 college students from 10 different majors corresponding to the subjects of our slides. We required annotators to have a GPA above 3.8 to ensure their expertise. To facilitate the annotation process, we build an online platform (See Appendix). Each annotator will get slides that match with their major. The platform displays each slide page alongside the corresponding script segment generated by the SFT model. We ask annotators to check and revise each page's script for the following error types: factual errors, missing information, relevance to the image, coherence of sentences, and repetition of words. After

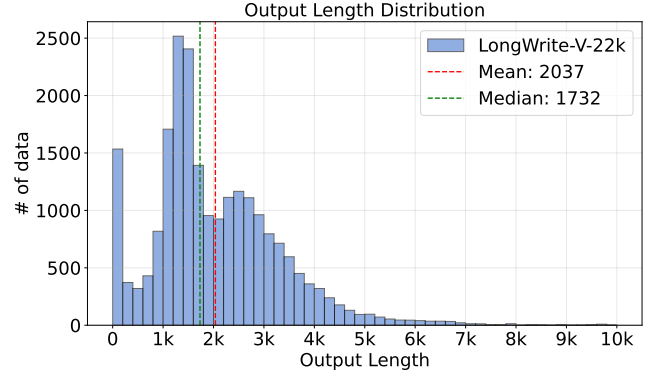


Figure 5: Output length statistics of LongWrite-V-22k.

completing the annotation of a slide, our authors will review the annotation quality. We get 72 valid scripts with human corrections.

3.3 Training

Supervised Fine-tuning. We conduct model training based on two open-source VLMs with different parameter sizes: Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-72B-Instruct [33]. We choose Qwen2.5-VL series as base model because they support a context window of 32k tokens. By resizing the input image's width and height to 280x280, the Qwen2.5-VL models can process up to 30 images. As shown in Figure 5, the output length in LongWrite-V-22k are distributed between 0 and 10k with two peaks around 0 and 1.5k. The peak at 0 indicates some short output data is mixed in the LongWrite-V-22k, which are mainly the results of those simple instructions. To get a better length distribution, we sample 10k data from LongWrite-V-22k with an average output length of 2.8k as training data. We then fine-tune the two models for 3 epochs with a learning rate of 1e-5 for Qwen2.5-VL-7B-Instruct and 7e-6 for Qwen2.5-VL-72B-Instruct, resulting in two models: LongWriter-V-7B and LongWriter-V-72B.

Iterative Direct Preference Optimization. After SFT phase, DPO [28] is a widely-used method to optimize VLM's output quality, which learns from a dataset of preference pairs $\mathcal{D} = \{(v, x, y_w, y_l)\}$, where the winning output y_w is preferred over the losing output y_l given the same visual input v and text input x . The optimization objective of DPO is to maximize the difference between likelihood of preference pairs:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(v, x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(\beta \log \frac{\pi_\theta(y_w|v, x)}{\pi_{\text{ref}}(y_w|v, x)} - \beta \log \frac{\pi_\theta(y_l|v, x)}{\pi_{\text{ref}}(y_l|v, x)})] \quad (4)$$

In our annotation process, v represents the images of a slide, x is the instruction for generating scripts, y_l is the flawed output script of VLM and y_w is the slide's lecture after human revision. However, collecting human feedback on long output is very time-consuming and expensive. As mentioned in Section 3.2.1, we gather 72 preference pairs on the scripts, which costs one week and around 1,000 \$ to finish. To make most use of these data, we propose to iteratively learn the fine-grained human correctional feedback on the long output. As the $y_w = \{y_w^i, i = 1, \dots, N\}$ is a revised script for

Table 1: Baseline models' context window lengths and number of max output tokens.

Model name	Context window	Max output tokens
<i>Large Language Models</i>		
GLM-4-9B-chat [11]	128,000 tokens	-
Mistral-Large-Instruct [16]	128,000 tokens	-
Deepseek-r1 [12]	64,000 tokens	8,000 tokens
<i>Vision Language Models</i>		
MiniCPM-V2.6 [44]	32,000 tokens	-
Qwen2.5-VL-7B [33]	32,000 tokens	-
Qwen2.5-VL-72B [33]	32,000 tokens	-
Claude 3 Opus [2]	200,000 tokens	4,096 tokens
Gemini-1.5-pro [32]	2,000,000 tokens	8,192 tokens
GPT-4o [26]	128,000 tokens	8,192 tokens

an N page slide, we increasingly view each page's script y_w^i as a winning segment over the flawed script:

$$\mathcal{L}_{\text{IterDPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(v, x, y_w, y_l) \sim \mathcal{D}} \sum_{i=1}^N \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w^{\leq i} | v_{\leq i}, x)}{\pi_{\text{ref}}(y_w^{\leq i} | v_{\leq i}, x)} - \beta \log \frac{\pi_\theta(y_l^{\leq i} | v_{\leq i}, x)}{\pi_{\text{ref}}(y_l^{\leq i} | v_{\leq i}, x)} \right) \right] \quad (5)$$

where $y_w^{\leq i}, y_l^{\leq i}$ is the revised and unrevised scripts until page i , and $v_{\leq i}$ are the corresponding images. We view $y_w^{\leq i}$ as a new winning response over the flawed output $y_l^{\leq i}$, this can help VLM learn the fine-grained feedback on the long output and extend the number preference pairs for N times. In this way, we get 1,477 iterative pairs for training. Apart from human feedback, we also utilize AI feedback by employing the gpt4o as the reward model. Following RLAI [47], we sample responses from the SFT model for 1,367 long-output instructions and use GPT-4o for assigning length and quality scores for the responses to construct preference pairs. Our final DPO model is trained with 2,844 mixed preference pairs,

4 Experiments

4.1 Experimental Setup

Metric. Following Bai et al. [3], we evaluate the VLM's output length and quality using two metrics: S_l and S_q . S_l is the output score that measures how close that the VLM's output length l_v is to the required length l_r . We use the number of words to count the output length. The length score S_l is defined as:

$$S_l = \begin{cases} 100 \cdot \max \left(0, 1 - \frac{(l_v/l_r - 1)}{3} \right) & \text{if } l_v > l_r, \\ 100 \cdot \max \left(0, 1 - \frac{(l_r/l_v - 1)}{2} \right) & \text{if } l_v \leq l_r. \end{cases} \quad (6)$$

We also use gpt-4o-2024-08-06 to assign the quality score S_q for six aspects: Relevance, Accuracy, Coherence, Clarity, Breadth and Depth, and Reading Experience. We list the scoring prompt in Appendix. Note that we have asked gpt-4o not to take the output length into account so that the quality score is independent with the length score. The overall score $\bar{S}$ is the mean of S_l and S_q .

Baselines. We evaluate 3 proprietary VLMs, 3 open-source VLMs and 4 LLMs on MMLongBench-Write (model details about models are listed in Table 1). Given that LLMs can also process visual

instructions via reading the image caption [25], we first use gpt-4o to describe the input images and then feed the caption and writing instruction to the LLM.

Implementation Details. The SFT phase for LongWriter-V-7B on the LongWrite-V-22k dataset took approximately six hours using 8 A800 GPUs. For the LongWriter-V-72B model, the SFT process required 72 hours with the same GPU configuration. The Direct Preference Optimization (DPO) process for LongWriter-V-7B, using 2,844 mixed preference pairs, completed in 1.5 hours.

4.2 Main Results

We report the performance of baselines and our trained models in Table 2. To study the effective output length of models, we divide the MMLongBench-Write benchmark into four subsets based on the instruction's required word count: 0-1500 words, 1500-2000 words, 2000-3000 words, and over 3000 words. The highest length and quality scores for each subset among models are in bold. We have three observations on the results: (1) Most existing models struggle to satisfy the length requirement over 2000 words, while LongWriter-V models can generate enough words for such instructions. By checking the length score S_l across different length intervals, we find that most models perform poorly on the [2000, 3000) range, with their S_l below 70. In contrast, our LongWriter-V models can generate outputs with effective length and high quality even on the range of [3000, 4000). (2) The scaling law effect on our benchmark is striking: smaller models like Qwen2.5-VL-7B-Instruct perform poorly in our evaluation with an overall score of 54.4, while its larger counterpart Qwen2.5-VL-72B-Instruct achieves a notably higher score of 83.3. Besides, after training the two VLMs on our LongWrite-V-22k data, both models improve significantly on long generation. The performance gap between the two sizes' models is narrowed after SFT (LongWriter-V-7B's 81.8 vs. LongWriter-V-72B's 84.9). (3) DPO can improve both the VLM's output quality and the ability to follow the length requirements of long generation. LongWriter-V-7B-DPO, which is the model trained on LongWriter-V-7B with 2,844 preference pairs, achieves improvement on both S_l (+3.7) and S_q (+1.8), showing that DPO is effective for boosting the long generation capabilities of VLMs.

4.3 Human Evaluation

As the quality score S_q is assigned by the GPT-4o automatically, the evaluation results may have bias as LLM tends to favor the responses generated by itself [17, 36]. To get a more fair quality comparison for the models, we conduct a human evaluation to capture the actual human preferences on model responses. Specifically, we select responses from four models: the three models trained by us and the GPT-4o-2024-08-06 baseline. We ask two human annotators to vote for their preferred response between two selected models on the 120 responses of MMLongBench-Write. For each annotator, we collect 720 votes and calculate the average win rate among models using two annotators' feedback.

The results are shown in Figure 6, where we surprisingly find that two of our trained models receive more votes from humans in the comparison with the GPT-4o-2024-08-06 baseline. While in the automatic quality score comparison, the two models also surpass the GPT-4o on the quality score. This indicates that our trained

Table 2: Evaluation results (%) on MMLongBench-Write. Note that LLMs are tested with input images transformed into captions. We report scores on different subsets of the benchmark, where [0,1000] means the expected output length falls within 0 to 1000 tokens. $\bar{S}$, S_l , S_q is the overall score, length score and quality score respectively.

Model	Overall			[0,1500)		[1500,2000)		[2000,3000)		[3000,4000)	
	$\bar{S}$	S_l	S_q	S_l	S_q	S_l	S_q	S_l	S_q	S_l	S_q
<i>Caption + LLMs</i>											
GLM-4-9B-Chat	71.3	62.0	80.6	87.9	72.2	65.7	82.4	44.7	76.7	24.2	93.5
GPT-4o-2024-08-06	77.1	66.6	87.5	86.7	81.2	68.9	88.3	58.7	85.8	33.5	97.2
Mistral-Large-Instruct-2407	78.9	69.6	88.2	89.7	84.7	70.9	89.9	58.4	83.0	47.2	94.9
DeepSeek-R1	82.4	70.3	94.5	87.2	92.4	73.4	95.7	59.8	92.0	38.1	95.8
<i>Open-source VLMs</i>											
MiniCPM-V2.6	54.1	30.3	77.8	56.1	68.9	31.3	81.7	15.0	69.4	4.5	86.1
Qwen2.5-VL-7B-Instruct	54.4	45.3	63.5	62.9	51.1	46.6	70.5	37.6	50.6	16.1	67.6
Qwen2.5-VL-72B-Instruct	83.3	79.9	86.7	80.0	78.4	84.5	90.3	71.6	79.7	65.3	91.7
<i>Proprietary VLMs</i>											
Claude-3-Opus-20240229	61.7	41.5	82.0	52.0	64.7	42.8	87.5	36.1	74.6	23.3	89.8
GPT-4o-2024-08-06	62.7	42.7	82.6	86.6	91.2	37.7	83.1	34.2	71.6	14.2	88.4
Gemini-1.5-Pro	83.0	74.8	91.2	88.7	93.0	78.1	91.8	62.2	86.2	50.5	95.4
<i>Our trained VLMs</i>											
LongWriter-V-7B	81.8	82.5	81.1	63.3	72.8	87.8	86.4	81.2	69.2	86.8	87.5
LongWriter-V-7B-DPO	84.6	86.2	82.9	69.5	82.5	90.5	86.9	87.1	69.0	87.4	85.2
LongWriter-V-72B	84.9	84.3	85.5	73.2	83.3	86.2	89.3	88.4	75.8	81.4	85.2

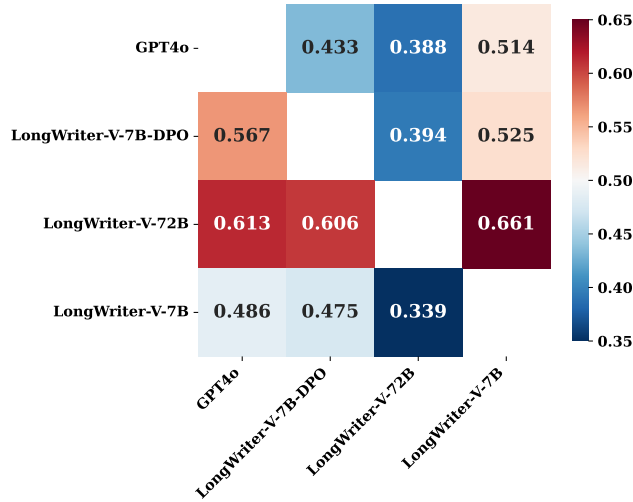


Figure 6: Human evaluation results on MMLongBench-Write, where each block of the matrix represents the model of the row’s win rate over the model of the column. The win rate is voted by two annotators.

models have gained some advantages over the GPT-4o baseline in the human preference, which is consistent with the automatic evaluation on the quality score of responses. Besides, we also display a case in Figure 7 to compare our tested models.

4.4 Ablation Study

We conduct ablation experiments on both the SFT and DPO process of LongWriter-V models. For the LongWriter-V-7B model trained on LongWrite-V-22k data, we control the three data sources of LongWrite-V-22k to observe how they contribute to the final performance of the SFT model. We run the SFT process on Qwen2.5-VL-7B-Instruct without (w/o) single-image, multi-image or back-translation data respectively and evaluate the trained models on MMLongBench-Write. As shown in Table 3, removing any of these data sources may lead to a decline in the overall score, where multi-image data is the most essential one, causing a decrease of 15.3 overall score. These results indicate that these sources are useful for training long output VLMs.

To explore the effectiveness of our DPO strategy over the small size preference data on long-output VLM alignment, we run the DPO process without those extra pairs extended by the iterative strategy. Results in Table 3 demonstrates that the model gains +1.1 length score but -1.1 quality score and -2.5 PPT task score over the DPO model with full data, which means the extended data is useful for the generation quality and the PPT script task. To examine the effectiveness of mixing AI preference pairs, we then train the SFT model with the human revised preference pairs only, resulting in a even worse performance (-1.1 overall score against the SFT model). This suggests that incorporating AI-generated pairs can improve model performance by providing additional training signals.

5 Related Work

Recent advancements in Vision-Language Models have focused on enhancing their ability to process long-context inputs [6, 10, 18].

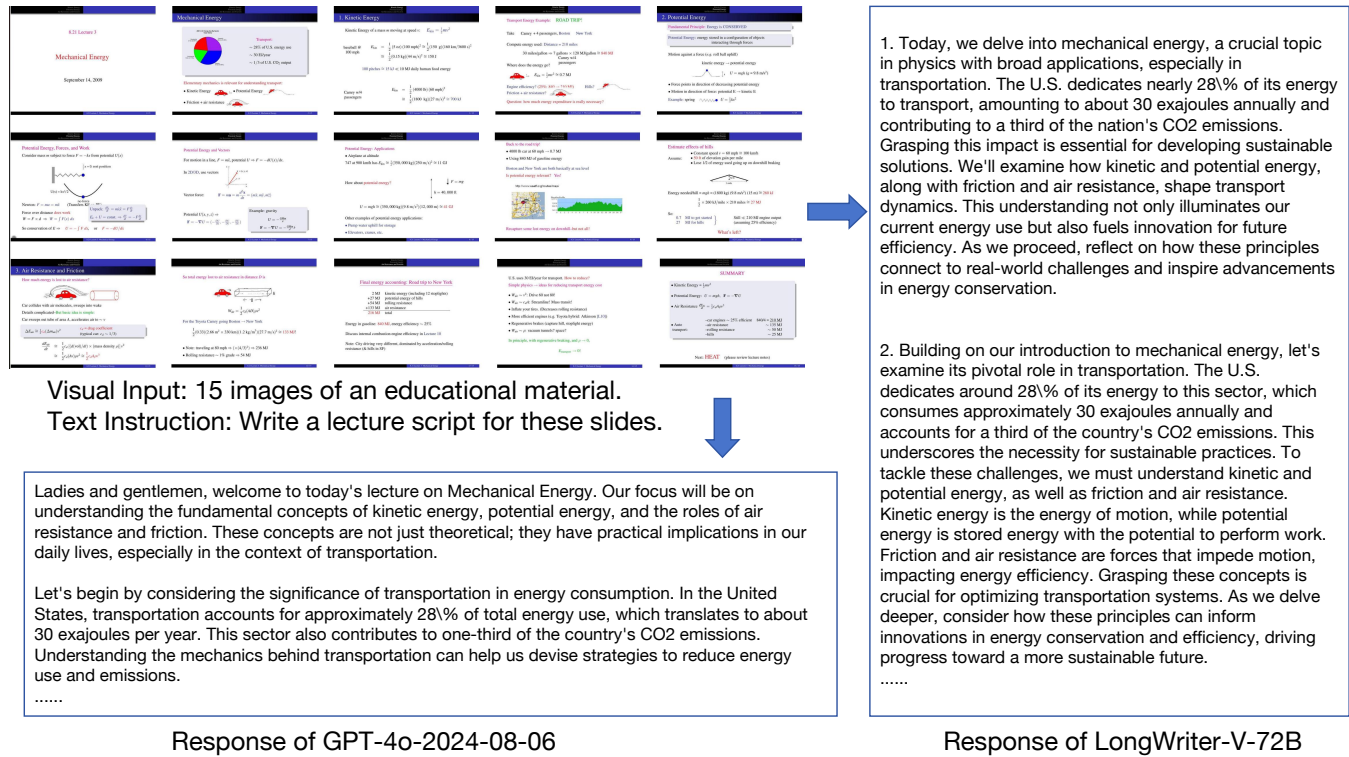


Figure 7: A case study for GPT-4o-2024-08-06 and LongWriter-V-72B on the lecture script generation task.

There are abundant benchmarks and datasets that designed for multimodal long context understanding including MMLongBench-Doc [25], LongDocURL [8], LongViTu [40], ShareGPT4Video [5], LongVideoBench [39] and LVBench [38]. However, the long-output generation abilities of VLMs have been less explored. In our work, we find that current VLMs struggle to generate an output with over 1000 tokens, which is much shorter than their max input context length (>16,000 tokens) [35]. To fill this gap, we explore how to extend the maximum output length of VLMs.

Although we show that supervised fine-tuning can align VLMs with user's instructions on length requirements, it is also important to improve the quality of long output [42]. Previous works mainly focus on how to improve VLMs' generation quality on short output tasks via post training methods such as RLHF-V [46], RLAI-F-V [47], POVID [50] and MIA-DPO [23]. However, none of these methods have explored how to effectively use human correctional feedback on long output for aligning VLMs. We propose to iteratively use each segment of the revised long output as the preferred response, which extends the number of preference pairs and successfully improves the long generation quality of VLM.

6 Conclusion

Our work introduces MMLongBench-Write, a comprehensive benchmark for evaluating long-generation tasks with visual inputs, and LongWriter-V-22k, a novel supervised fine-tuning dataset designed to enhance the long-output capabilities of VLMs. We demonstrate that the SFT model trained on LongWrite-V-22k can generate long

Table 3: Scores (%) on MMLongBench-Write for models trained under different conditions, where S , S_l and S_q is the overall, length and quality score on all tasks and S_{PPT} is the overall score on the PPT script task.

Model	S	S_l	S_q	S_{PPT}
LongWriter-V-7B	81.8	82.5	81.1	83.1
w/o single-image data	79.6	79.5	79.6	83.4
w/o multi-image data	66.5	60.3	72.7	29.3
w/o backtranslation	80.7	80.0	81.3	82.4
LongWriter-V-7B-DPO	84.6	86.3	82.9	85.8
w/o iterative pairs	84.6	87.4	81.8	83.3
w/o 1.4k gpt4o feedback	80.7	78.7	82.7	71.7

outputs with high fidelity, outperforming existing models in terms of both length and quality. Furthermore, our proposed IterDPO method effectively leverages human feedback to improve the fidelity of long outputs, addressing issues such as hallucination. Future research may explore more training strategies and larger datasets to further push the boundaries of long-output generation in VLMs. Moreover, reinforcement learning methods [14] that guide the VLMs to generate long outputs [41] with high quality and fidelity are also worth exploring.

Acknowledgments

This work is supported by Beijing Natural Science Foundation (L243006) and National Natural Science Foundation of China (62476150, 62277033 and 62407027). Thanks the support from National Engineering Laboratory for Cyberlearning and Intelligent Technology. This work is also supported by a grant from the Institute for Guo Qiang, Tsinghua University.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.
- [2] Anthropic. 2024. Anthropic: Introducing the next generation of Claude. <https://www.anthropic.com/news/claude-3-family>
- [3] 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* (2024).
- [4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharept4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*. Springer, 370–387.
- [5] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. 2024. Sharept4videio: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325* (2024).
- [6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.
- [7] Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E Gonzalez, and Wei-Lin Chiang. 2024. VisionArena: 230K Real World User-VLM Conversations with Preference Labels. *arXiv preprint arXiv:2412.08687* (2024).
- [8] Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. 2024. LongDocURL: a Comprehensive Multimodal Long Document Benchmark Integrating Understanding, Reasoning, and Locating. *arXiv preprint arXiv:2412.18424* (2024).
- [9] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14303–14312.
- [10] 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* (2024).
- [11] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadao Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuntao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793* (2024).
- [12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
- [13] Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence* 7 (2024), 1430984.
- [14] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. 2025. GLM-4.1 V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. *arXiv preprint arXiv:2507.01006* (2025).
- [15] Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics* 11 (2023), 565–581.
- [16] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
- [17] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llm-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579* (2024).
- [18] Mukai Li, Lei Li, Shanshan Gong, and Qi Liu. 2024. GIRAFFE: Design Choices for Extending the Context Length of Visual Language Models. *arXiv preprint arXiv:2412.12735* (2024).
- [19] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259* (2023).
- [20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.
- [21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*. Springer, 38–55.
- [23] Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637* (2024).
- [24] Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. 2024. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840* (2024).
- [25] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523* (2024).
- [26] OpenAI. 2024. OpenAI: Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>
- [27] Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following for long-form text generation. *arXiv preprint arXiv:2406.19371* (2024).
- [28] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).
- [29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.
- [30] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. 2024. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485* (2024).
- [31] Irma Taavitsainen and Päivi Pahta. 2000. Conventions of professional writing: The medical case report in a historical perspective. *Journal of English Linguistics* 28, 1 (2000), 60–76.
- [32] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
- [33] Qwen Team. 2025. Qwen2.5-VL. <https://qwenlm.github.io/blog/qwen2.5-vl/>
- [34] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100* (2022).
- [35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
- [36] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghui Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926* (2023).
- [37] Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuoxu Xu, Jihong Dai, et al. 2024. Weaver: Foundation models for creative writing. *arXiv preprint arXiv:2401.17268* (2024).
- [38] Weihang Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. 2024. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035* (2024).
- [39] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv*

- preprint arXiv:2407.15754* (2024).
- [40] Rujie Wu, Xiaojian Ma, Hai Ci, Yue Fan, Yuxuan Wang, Haozhe Zhao, Qing Li, and Yizhou Wang. 2025. LongViTU: Instruction Tuning for Long-Form Video Understanding. *arXiv preprint arXiv:2501.05037* (2025).
 - [41] Yuhao Wu, Yushi Bai, Zhiqiang Hu, Roy Ka-Wei Lee, and Juanzi Li. 2025. LongWriter-Zero: Mastering Ultra-Long Text Generation via Reinforcement Learning. *arXiv preprint arXiv:2506.18841* (2025).
 - [42] Yuhao Wu, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. 2024. LongGen-Bench: Benchmarking Long-Form Generation in Long Context LLMs. *arXiv preprint arXiv:2409.02076* (2024).
 - [43] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. 2024. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188* (2024).
 - [44] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* (2024).
 - [45] Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, Wenzheng Feng, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, et al. 2020. MOOCCube: A large-scale data repository for NLP applications in MOOCs. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3135–3142.
 - [46] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. RLhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13807–13816.
 - [47] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024. RLai-f-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220* (2024).
 - [48] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
 - [49] Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2025. PPTAgent: Generating and Evaluating Presentations Beyond Text-to-Slides. *arXiv:2501.03936 [cs.AI]* <https://arxiv.org/abs/2501.03936>
 - [50] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411* (2024).