Rethinking the Text-Vision Reasoning Imbalance in MLLMs through the Lens of Training Recipes

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have demonstrated strong capabilities on vision-and-language tasks. However, recent findings reveal an imbalance in their reasoning capabilities across visual and textual modalities. Specifically, current MLLMs often over-rely on textual cues while under-attending to visual content, resulting in suboptimal performance on tasks that require genuine visual reasoning. We refer to this phenomenon as the *modality gap*, defined as the performance disparity between text-centric and vision-centric inputs. In this paper, we analyze the modality gap through the lens of training recipes. We first show that existing training recipes tend to amplify this gap. Then, we systematically explore strategies to bridge it from two complementary perspectives: data and loss design. Our findings provide insights into developing training recipes that mitigate the modality gap and promote a more balanced multimodal reasoning.

1 Introduction

Multimodal large language models (MLLMs) have shown exceptional reasoning capabilities on complex tasks that require multimodal reasoning. However, recent studies (Zhang et al., 2024; Li et al., 2025) reveal a reasoning imbalance: these models often rely heavily on textual cues while underexploiting visual information when generating answers. This over-reliance on text leads to suboptimal results on tasks that require genuine visual reasoning. We refer to this phenomenon as the *modality gap*. As exemplified in Figure 1, when critical information present in the visual modality is removed from the text, MLLMs fail to answer questions that could have been correctly answered when the full text was provided, highlighting their insufficient visual reasoning.

To understand the origin of this imbalance, we focus on the training recipes of current MLLMs.

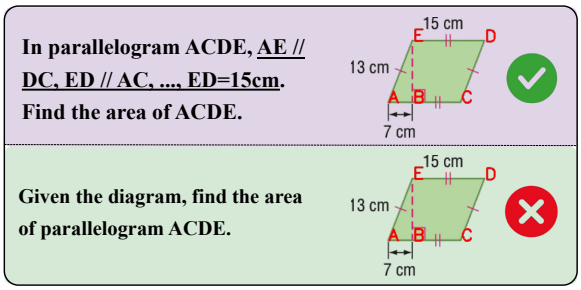


Figure 1: Current MLLMs exhibit an imbalance between visual and textual reasoning. When information present in the visual modality is removed from the text, the MLLM fails to answer the question.

An important observation is that many training samples contain overlapping information across the textual and visual modalities. In such cases, it may be easier for MLLMs to rely on the already complete textual information rather than engage in visual reasoning. We hypothesize that this training process largely contributes to the modality gap. Our preliminary evidence supports this view: under standard training setups, the gap between text- and vision-centric performance widens over time, underscoring the need for more balanced training strategies.

Building on these insights, our goal is to identify improved training recipes for MLLMs that jointly ① ensure the effective use of visual information and ② maintain or enhance overall reasoning ability in the targeted domains. We approach this problem from two perspectives: *data* and *loss*. From the data perspective, we consider vision-centric and text-centric data in supporting balanced multimodal reasoning and explore how simple data mixing and carefully designed curriculum strategies can leverage the strengths of both modalities. From the loss perspective, we propose a KL-based self-distillation objective that transfers reasoning competence from full-text to partial-text inputs, preserving general reasoning performance while strengthening visual grounding. Our key contribu-

tions are as follows:

- We establish a diagnostic that reveals a consistent discrepancy between text-centric and vision-centric performance across a range of public and private MLLMs at different scales. Furthermore, we show that current training setups *widen* this discrepancy, underscoring the need for targeted recipes beyond naïve RL.

- We propose improved RL training recipes from the perspectives of *data* and *loss*, designed to preserve reasoning competence in the textual modality while reducing the reasoning gap across modalities.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) have emerged as powerful tools that integrate visual and textual information to perform a wide range of tasks, including image captioning, visual question answering, and geometric reasoning. Notable MLLMs include Qwen2.5-VL series (Bai et al., 2025; Wang et al., 2024; Bai et al., 2023), VL-Rethinker series (Wang et al., 2025a), MiniCPM series (Yao et al., 2024), InternVL3.5 series (Wang et al., 2025b), Kimi-VL series (Team et al., 2025b), Gemma series (Team et al., 2025a), GPT-5 (OpenAI, 2025) and Gemini (Comanici et al., 2025). These models typically employ a combination of pre-trained vision encoders and large language models, fine-tuned on multimodal datasets to enhance their understanding and generation capabilities across both modalities.

2.2 Visual Reasoning in MLLMs

Visual reasoning is a critical capability for MLLMs, enabling them to interpret and reason about visual content in conjunction with textual information. Recent studies like MathVerse (Zhang et al., 2024) have highlighted the challenges MLLMs face in effectively utilizing visual information, often defaulting to text-based cues. This has led to the identification of the modality gap, where models perform significantly better on text-centric tasks compared to vision-centric ones.

To address this issue, various approaches have been proposed, including specialized training datasets (Liu et al., 2024; Li et al., 2024; Gao et al., 2023), model architecture design (Lu et al., 2024; Bigverdi et al., 2025), and loss functions that encourage visual attention (Luo et al., 2024; Li et al., 2025; Wang et al., 2025c). In this paper, we build upon these foundations by exploring RL-based methods to enhance visual reasoning while mitigating the modality gap.

Model	Dataset	Text	Vision	Avg	Gap
Qwen2.5-VL 3B	PGPS9K	*23.97*	*18.12*	*21.05*	***5.85***
	MathVerse	*35.68*	*28.66*	*31.47*	*7.02*
Qwen2.5-VL 7B	PGPS9K	*37.75*	*29.98*	*33.87*	*7.77*
	MathVerse	*55.01*	*45.76*	*51.10*	*9.25*
MiniCPM-V-4	PGPS9K	*34.70*	*30.20*	*32.45*	*4.50*
	MathVerse	*44.35*	*37.21*	*40.07*	*7.14*
Gemma-3-4b-it	PGPS9K	*40.50*	*26.12*	*35.59*	*14.38*
	MathVerse	*42.36*	*33.50*	*37.05*	*8.86*
Kimi-VL-A3B	PGPS9K	*40.57*	*31.32*	*35.95*	*9.25*
	MathVerse	*57.91*	*48.27*	*51.23*	*9.64*
VL-Rethinker 7B	PGPS9K	*40.45*	*36.05*	*38.25*	*4.40*
	MathVerse	*65.42*	*57.28*	*60.53*	*8.14*
InternVL3.5 8B	PGPS9K	*52.18*	*39.65*	*45.92*	*12.53*
	MathVerse	*66.68*	*54.19*	*59.19*	*12.49*
Qwen3-VL	PGPS9K	*69.70*	*66.99*	*68.35*	*2.71*
	MathVerse	*64.06*	*60.89*	*61.67*	***3.17***
GPT-5[1]	PGPS9K	***94.00***	***80.00***	***87.00***	*14.00*
	MathVerse	***76.67***	*63.33*	*70.00*	*13.34*
Gemini 2.5 Flash[1]	PGPS9K	*92.00*	*74.00*	*83.00*	*18.00*
	MathVerse	*86.96*	***77.78***	***82.00***	*9.18*

Table 1: Base model performance.

3 Modality Gap in MLLMs

We begin by quantifying the modality gap across a range of open-source and commercial MLLMs. To illustrate this gap, we consider two kinds of data:
• $\mathcal{D}_1$: **Text-centric.** All necessary information is contained within the provided text, and the MLLM can solve the problem through textual reasoning.
• $\mathcal{D}_2$: **Vision-centric.** Some necessary information is present in the image but not in the text, requiring the MLLM to perform visual reasoning to successfully solve the problem.

To construct $\mathcal{D}_1$ and $\mathcal{D}_2$, we draw upon two challenging visual reasoning datasets: PGPS9K (Zhang et al., 2023) and MathVerse (Zhang et al., 2024). In PGPS9K, each question consists of a textual condition and a question statement, accompanied by a fully annotated figure that specifies entities and their relations. Accordingly, we define $\mathcal{D}_1$ as the setting where both the image and text provide complete information, and $\mathcal{D}_2$ as the setting where information present in the image has been removed from the text. For MathVerse, following prior work, we define $\mathcal{D}_1$ and $\mathcal{D}_2$ subsets to focus respectively on text (*Text-Dominant* and *Text-Lite* subsets) and vision (*Vision-Intensive*, *Vision-Dominant*, and

[1]Due to API costs, the results are evaluated on a subset of 50 test samples.

Model	Text-centric	Vision-centric	Avg
Qwen2.5-VL 3B	*23.97*	*18.12*	*21.05*
Qwen2.5-VL 3B $\mathcal{D}_1$	***62.08***	*44.80*	***52.49***
Qwen2.5-VL 3B $\mathcal{D}_2$	*51.18*	***50.50***	*50.84*
Qwen2.5-VL 7B	*37.75*	*29.98*	*33.87*
Qwen2.5-VL 7B $\mathcal{D}_1$	***74.22***	*53.77*	***64.00***
Qwen2.5-VL 7B $\mathcal{D}_2$	*63.42*	***60.40***	*61.91*

Table 2: Standard RL training on PGPS9K Results.

Vision-Only subsets) reasoning capabilities. Further details of the datasets are provided in Appendix A.

Metrics. We report the **text-centric** and **vision-centric** performance measured on $\mathcal{D}_1$ and $\mathcal{D}_2$. In addition, we report the **overall** performance as the average accuracy across $\mathcal{D}_1$ and $\mathcal{D}_2$.

Direct Inference Results. We begin by evaluating a series of off-the-shelf MLLMs. The results are summarized in Table 1. Across both the PGPS9K and MathVerse datasets, we observe a consistent modality gap: text-centric performance is consistently higher than vision-centric performance across various open-source and commercial models of different sizes. Moreover, stronger MLLMs tend to exhibit a larger performance gap. This discrepancy underscores the need for targeted strategies to enhance the visual reasoning capabilities of MLLMs.

Effect of Standard RL Training. Next, we explore how standard training influences the modality gap. In this experiment, we apply DAPO (Yu et al., 2025) to fine-tune Qwen2.5-VL (3B and 7B) under both $\mathcal{D}_1$ and $\mathcal{D}_2$ settings from the PGPS9K training set. Note that all figures in $\mathcal{D}_1$ and $\mathcal{D}_2$ have their entities, relations, and other geometric properties explicitly annotated. Thus, the model can always obtain complete information related to the question from the image. As shown in Table 2, training on $\mathcal{D}_1$ primarily improves text-centric performance but enlarges the modality gap as training progresses, whereas training on $\mathcal{D}_2$ strengthens vision-centric performance and narrows the gap, though at the expense of overall accuracy.

Moreover, as shown in Figure 2, during standard training on $\mathcal{D}_1$, the modality gap progressively increases with training steps. These observations indicate that the standard training recipe is insufficient to resolve the modality gap in MLLMs, highlighting the need for more nuanced training strategies.

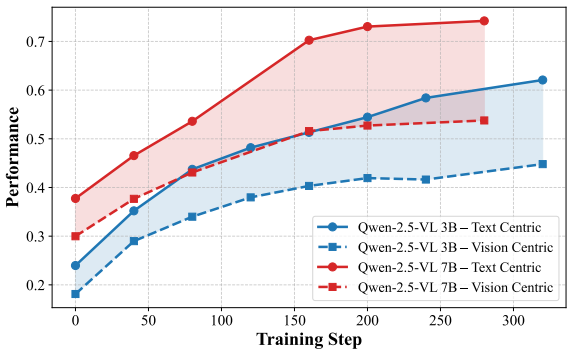


Figure 2: Standard training recipe widens modality gap.

4 Mitigating Modality Gap

In this section, we investigate an improved RL training recipe from two complementary perspectives to enhance the visual reasoning ability of MLLMs:

• **Data.** We explore two training strategies: ① *mixed training*, which combines $\mathcal{D}_1$ (full-text) and $\mathcal{D}_2$ (partial-text) samples to expose models to both text- and vision-centric inputs; and ② *curriculum training*, which first trains on $\mathcal{D}_1$ to consolidate reasoning under textual guidance, and then shifts to $\mathcal{D}_2$ to strengthen image-based reasoning and reduce shortcut reliance.

• **Loss.** We introduce a KL-based self-distillation loss to align the model's output distribution on $\mathcal{D}_2$ with that on $\mathcal{D}_1$, thereby preserving core reasoning ability while enhancing visual understanding.

4.1 Data Perspective

Implementation. We compare *mixed training* and *curriculum training*, matching the total training budget (details in Appendix E). Curriculum training splits steps evenly: Stage 1 on $\mathcal{D}_1$, followed by Stage 2 on $\mathcal{D}_2$.

Result. As summarized in Table 3, curriculum training generally matches or surpasses mixed-data training in both in-distribution (PGPS9K) and out-of-distribution (MathVerse) evaluations. Intuitively, Stage 1 on $\mathcal{D}_1$ consolidates general reasoning and solution formatting under rich textual guidance; Stage 2 on $\mathcal{D}_2$ then compels stronger visual grounding. This two-stage approach effectively improves both text-centric and vision-centric performance.

4.2 Loss Perspective

Implementation. We introduce a *contrastive self-distillation KL loss* to transfer reasoning from full-text ($\mathcal{D}_1$) to partial-text ($\mathcal{D}_2$) inputs. Given paired prompts $(x^{(1)}, x^{(2)})$, and a correct response

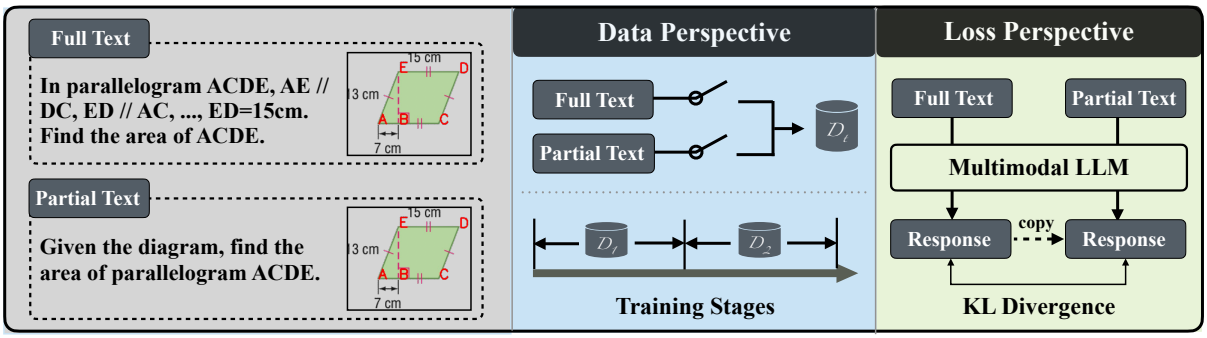


Figure 3: We consider two types of data: ❶ both text and image contain complete information, referred to as *full text*; and ❷ the text omits information already present in the image, referred to as *partial text*. We then analyze better training recipe from both data and loss perspectives.

Training Strategy	PGPS9K			MathVerse		
	Text	Vision	Avg	Text	Vision	Avg
Qwen2.5-VL 3B						
Mixed training	57.93	52.80	55.37	46.07	41.11	43.09
Curriculum Stage 1 ($\mathcal{D}_1$)	58.40	41.63	50.02	**49.37**	43.17	45.65
Curriculum Stage 2 ($\mathcal{D}_1 \rightarrow \mathcal{D}_2$)	**59.78**	**54.00**	**56.89**	49.02	**43.69**	**45.82**
Qwen2.5-VL 7B						
Mixed training	70.10	65.65	67.88	54.93	48.33	50.97
Curriculum Stage 1 ($\mathcal{D}_1$)	**73.05**	52.72	62.89	49.95	44.42	46.63
Curriculum Stage 2 ($\mathcal{D}_1 \rightarrow \mathcal{D}_2$)	70.60	**66.30**	**68.45**	**56.95**	**50.60**	**53.14**

Table 3: Data mixing and curriculum training results

Training Strategy	PGPS9K			MathVerse		
	Text	Vision	Avg	Text	Vision	Avg
Qwen2.5-VL 3B						
Plain RL on $\mathcal{D}_1$	58.40	41.63	50.02	49.37	43.17	45.65
w/ KL	58.10	45.47	51.79	**49.57**	**43.67**	**46.03**
w/ KL + Curriculum	**61.22**	**55.27**	**58.25**	47.08	42.23	44.17
Qwen2.5-VL 7B						
Plain RL on $\mathcal{D}_1$	73.05	52.72	62.89	49.95	44.42	46.63
w/ KL	73.28	54.30	63.79	**57.00**	**48.79**	**52.07**
w/ KL + Curriculum	**73.42**	**67.87**	**70.65**	53.76	48.55	50.63

Table 4: Loss perspective results.

$\hat{\mathbf{y}}$ sampled from $\pi_\theta(\cdot \mid x^{(1)})$, we align the partial-text distribution $p_t := \pi_\theta(\cdot \mid \hat{y}_{<t}, x^{(2)})$ with the frozen full-text distribution $q_t := \mathrm{stopgrad}[\pi_\theta(\cdot \mid \hat{y}_{<t}, x^{(1)})]$ via a time-averaged forward KL:

$$\mathcal{L}_{\mathrm{cKL}}(\theta) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}\left(p_t \,\|\, q_t\right). \qquad (1)$$

The forward KL encourages the model's response distribution under partial-text inputs to *cover* the high-confidence region of its own distribution under full-text inputs. In practice, this KL loss is computed for all rollouts (without DAPO roll out batch group filtering) and added to the RL objective with weight $\alpha = 0.01$, providing a dense learning signal and helping maintain the overall training loss optimization process stable. After the contrastive KL loss has stabilized, the model is further fine-tuned on $\mathcal{D}_2$ to enhance its visual reasoning ability.

Result. We compare three training strategies in Table 4: ① Plain RL on $\mathcal{D}_1$, ② with KL, i.e., adding the contrastive KL loss, and ③ with KL + Curriculum, where the KL-trained model is subsequently fine-tuned on $\mathcal{D}_2$. From the in-distribution results on PGPS9K, both KL and KL + Curriculum consistently outperform the plain baseline, confirming that the KL term effectively transfers reasoning ability and stabilizes the optimization process. However, on the out-of-distribution dataset MathVerse, the improvements are less consistent, likely due to annotation and representation mismatches between the datasets. Specifically, PGPS9K provides explicit geometric cues, whereas MathVerse often omits such markings, weakening cross-domain transfer. We analyze this mismatch further in Appendix H. Overall, the KL loss enhances general reasoning ability, while the subsequent curriculum fine-tuning slightly degrades out-of-distribution performance, reflecting the impact of differing annotation styles across datasets. Please refer to Appendix G for more comparisons with baseline methods and Appendix F for additional ablation studies.

5 Conclusion

We present a systematic study on reducing the modality gap of MLLMs through reinforcement learning. Our experiments show that curriculum training effectively balances text-centric and vision-centric reasoning, and a KL-based self-distillation loss transfers reasoning competence from text-rich to vision-centric inputs. Together, these findings yield practical guidance: favor curriculum + contrastive KL to build MLLMs with stronger and more balanced visual reasoning capabilities.

Limitations

Our proposed training recipe relies on constructing paired text-centric and vision-centric data, which currently leverages the detailed annotations of the PGPS9K geometry dataset. While we demonstrate that this method enhances performance across both domain-specific and general multimodal benchmarks (as shown in Appendix G), extending this data construction strategy to broader, less structured VQA tasks remains a challenge. Future work will explore automated methods to generate such training pairs for diverse domains, further validating the scalability of our approach.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. 2025. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3836–3845.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

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

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Mingxiao Li, Na Su, Fang Qu, Zhizhou Zhong, Ziyang Chen, Yuan Li, Zhaopeng Tu, and Xiaolong Li. 2025. Vista: Enhancing vision-text alignment in mllms via cross-modal mutual information maximization. *arXiv preprint arXiv:2505.10917*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Xiaobo Xia, Tongliang Liu, and 1 others. 2024. Deem: Diffusion models serve as the eyes of large language models for image perception. *arXiv preprint arXiv:2405.15232*.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5. Accessed: 2025-10-04.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

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

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

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang

Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and 1 others. 2025c. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023. A multi-modal neural geometric solver with textual clauses parsed from diagram. *arXiv preprint arXiv:2302.11097*.

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

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024. Mmicl: Empowering vision-language model with multi-modal in-context learning. *Preprint*, arXiv:2309.07915.

A Dataset Details

PGPS9K. PGPS9K is a large-scale, human-annotated dataset containing over 9000 plain-geometry questions, split into 8000 training and 1000 test samples.

Each question comprises two components: a *textual condition* and a *question statement*. The textual condition fully specifies the geometric construction—listing entities such as points, lines, and circles and relations including parallelism, perpendicularity, and congruence—while the question statement queries a particular geometric property (e.g., the length of a segment or the measure of an angle).

All figures include explicit annotations of entities and relations, which we refer to as *full-condition images*. All questions are free-form and admit a unique numerical answer.

Based on these full-condition images, we define two dataset settings:

- $\mathcal{D}_1$: **Full-condition question + full-condition image**. The textual condition fully specifies the geometry, resembling text-centric setups in typical VQA or reasoning datasets.
- $\mathcal{D}_2$: **Question only + full-condition image**. The textual condition is omitted, requiring the model to infer the geometry directly from the image, resulting in a more vision-centric and challenging setting.

MathVerse. We adopt the open-source subset `testmini` of the MathVerse dataset as our out-of-distribution evaluation benchmark.

MathVerse contains two types of questions: multiple-choice and free-form questions, covering a broad range of visual-mathematical reasoning scenarios.

The subset used in our experiments includes 778 unique base questions, each instantiated into five variations: *Text Dominant*, *Text Lite*, *Vision Intensive*, *Vision Dominant*, and *Vision Only*, yielding a total of 3890 evaluation samples. These variations are designed to progressively reduce textual information while increasing dependence on visual cues, thus providing a systematic means of assessing the visual reasoning capability of multimodal large language models (MLLMs).

For evaluation, we report three metrics: ① the average accuracy across all five variations, ② the average accuracy on the three vision-centric variations (*Vision Intensive*, *Vision Dominant*, and *Vision Only*), and ③ the average accuracy on the two text-centric variations (*Text Dominant* and *Text Lite*).

B Evaluation

All results reported in this paper are obtained by sampling **four responses per question** (with a maximum response length of 4096 tokens) and averaging Pass@1 across the samples.

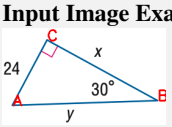
For PGPS9K, we extract the final numerical answer from each response using regular expressions

and compare it to the ground truth answer, which is also a number. A response is considered correct if the relative error is within 10^{-2}.

For MathVerse, we also extract the final numerical answer from each response using regular expressions. However, since MathVerse includes both multiple-choice and free-form questions, we evaluate them differently: for multiple-choice questions, a response is correct if the extracted answer matches the correct choice; for free-form questions, a response is correct if the relative error is within 5×10^{-2}.

C Complete Prompt

System Prompt
FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in `\boxed{<final answer>}`.
Input Image Example:



Prompt For Text-centric Task
In this problem, $CB \perp CA$ at C, $AC = 24$, $BC = x$, $AB = y$, and $m\angle CBA = 30°$. Based on these conditions, answer the question: Find y.
Prompt For Vision-centric Task
Based on the conditions in the image, answer the question: Find y.

D Dynamic sAmpling Policy Optimization(DAPO) Overview

DAPO(Yu et al., 2025) serves as the base RL training method in our pipeline. For clarity, relative to the standard GRPO framework, DAPO introduces the following modifications:

- **Clip-Higher.** DAPO decouples the lower and upper clipping ranges (-low,-high) of the importance-sampling ratios and enlarges the upper bound. This mitigates entropy collapse and improves exploration by allowing greater flexibility for low-probability exploration tokens during policy updates.

- **Dynamic sampling.** DAPO filters out prompts whose response groups are uniformly correct or uniformly incorrect, as these prompts do not generate meaningful gradients under GRPO. Since this filtering reduces the effective batch size, DAPO oversamples prompts that yield non-trivial advantages to maintain a sufficient number of gradient-contributing samples. This increases the density of useful training signals and improves the stability of GRPO-style updates.

- **Token-level policy gradient loss.** In vanilla GRPO, all token log-probability terms within a response are averaged first, and this averaged value becomes the sample's sole contribution to the update. Consequently, samples of different lengths receive equal weight, which dilutes the gradients of long chain-of-thought responses and overemphasizes short responses. DAPO eliminates this sample-level averaging and aggregates losses at the token level. Each token therefore contributes directly to the optimization objective, preventing long responses from being under-weighted and enabling finer-grained credit assignment. This produces more stable updates in long sequence RL and ensures that informative reasoning steps and error-inducing tokens are appropriately reflected in the gradient signal.

- **Overlong reward shaping.** To address instability caused by excessively long or truncated responses, DAPO applies a length-aware penalty known as Soft Overlong Punishment. When a response exceeds a predefined maximum length, DAPO introduces a punishment interval in which the penalty increases smoothly with response length. Shorter responses receive no penalty, while severely overlong responses receive a fixed maximum penalty. This penalty is added to the rule-based correctness reward to discourage unnecessarily long outputs while preserving valid reasoning content. This mechanism reduces reward noise from truncation and prevents models from exploiting longer outputs to manipulate rewards.

E Training Setup

All training for RL and ablations is conducted on the PGPS9K training set, and evaluation is performed on the PGPS9K test split and the MathVerse testmini subset. All reinforcement learning experiments are conducted with DAPO under the following configuration:

Clipping ratios. Lower and upper clipping thresholds are set to 0.2 and 0.28, with an additional coefficient $c = 10.0$ for actor-critic stability.

Model	Text-centric ($\mathcal{D}_1$)	Vision-centric ($\mathcal{D}_2$)	Text-only	Gap
Qwen3B Base Model	23.97	18.12	16.10	−2.02
Qwen3B Plain RL 240	56.50	**46.45**	46.82	0.37
Qwen3B Plain RL 320	**62.08**	44.80	**48.70**	3.90
Qwen7B Base Model	37.75	29.98	32.95	2.97
Qwen7B Plain RL 200	72.68	53.17	60.02	6.85
Qwen7B Plain RL 280	**74.22**	**53.77**	**62.10**	8.33

Table 5: Ablation results under a fixed training distribution. All models are trained exclusively on the Text-centric ($\mathcal{D}_1$) subset. The gap (Text-only − Vision-centric) increases consistently with additional RL steps, indicating that reinforcement learning disproportionately strengthens text-based reasoning.

Overlong responses. To handle long generations, we use a buffer length of 1024, enable buffer control, and apply a penalty factor of 1.0 when responses exceed this limit.

Training configuration. Batch size is 512 and mini-batch size is 128, with maximum prompt length of 1024 and maximum response length of 4096. The learning rate is fixed at 1×10^{-6}.

Stopping criterion. Unless otherwise noted, training is stopped once the DAPO parameter `num_gen_batches` reaches 10, which means that 10 rollout steps are required to accumulate one gradient update.

Models. We use Qwen2.5-VL 3B and 7B as our base models, which are open-source MLLMs with strong performance on visual reasoning tasks.

Computing Infrastructure. All experiments are conducted on 8 H100 GPUs with 80GB memory each. Each training run takes approximately 24 hours for Qwen2.5-VL 3B and 48 hours for Qwen2.5-VL 7B.

These settings are used consistently across all experiments to ensure comparability.

F Ablation on RL under a Fixed Training Distribution

To determine whether the observed modality gap is caused by RL training itself, rather than by a train–test distribution mismatch induced by splits of PGPS9K, we conduct a controlled evaluation in which the training distribution is strictly fixed. All models are trained exclusively on the **Text-centric** ($\mathcal{D}_1$) subset of PGPS9K, where both full textual descriptions and images are available. No modality-ablated subsets are used during training, and all RL updates are derived solely from this data.

After training, the same model checkpoints are evaluated under three inference settings: (1) **Text-centric** ($\mathcal{D}_1$), which includes both complete text and image inputs; (2) **Vision-centric** ($\mathcal{D}_2$), where textual descriptions are omitted, leaving only images and questions; and (3) **Text-only**, where images are withheld, providing only textual descriptions and questions. Since the training data distribution is identical across all models, performance differences across evaluation settings reflect changes in inference behavior rather than distribution mismatch.

Table 5 reports the results for Qwen models of different sizes and RL training steps. As RL progresses, performance improves consistently across all evaluation settings; however, the improvement is not uniform. Specifically, performance in the **Text-only** setting increases more rapidly than in the **Vision-centric** setting. For both 3B and 7B models, later-stage RL checkpoints exhibit higher accuracy in the Text-only setting compared to the Vision-centric setting, even though the models were never explicitly trained on text-only data.

This asymmetric improvement leads to a widening gap between text-only and vision-centric performance as RL steps increase. In contrast, base models exhibit a much smaller gap, sometimes performing comparably or even better in the Vision-centric setting. Because all models are trained on the same Text-centric ($\mathcal{D}_1$) dataset, these results indicate that reinforcement learning intrinsically biases optimization toward text-dominant reasoning, thereby amplifying the modality gap even when the training distribution is held constant.

G Comparison With Other Baseline Methods and General Benchmarks

To provide a direct comparison with existing baseline methods and assess the effectiveness of our proposed training strategy, we include two widely used baselines for multimodal reasoning. The first is an in-context learning baseline following MMICL (Zhao et al., 2024), where four multimodal examples are provided at inference time. The second is Perception-Aware Policy Optimization (PAPO) (Wang et al., 2025c), a reinforcement learning method designed to enhance multimodal perception during training. All models are trained exclusively on PGPS9K train set. All benchmarks discussed below are used solely for evaluation and comparison across different training strategies. All baselines are evaluated under the same inference protocol as our method.

We first compare these baselines on PGPS9K under both text and vision evaluation settings. Re-

Model	Text-centric	Vision-centric	Avg
Qwen 2.5 7B Base Model	37.75	29.98	33.87
Qwen 2.5 7B MMICL	37.18	30.70	33.94
Qwen 2.5 7B PAPO	45.87	37.70	41.79
Qwen 2.5 7B KL+Curriculum (Ours)	**73.42**	**67.87**	**70.65**
Qwen 2.5 3B Base Model	23.97	18.12	21.05
Qwen 2.5 3B MMICL	24.07	20.10	22.14
Qwen 2.5 3B PAPO	39.87	33.67	36.77
Qwen 2.5 3B KL+Curriculum (Ours)	**61.22**	**55.27**	**58.25**

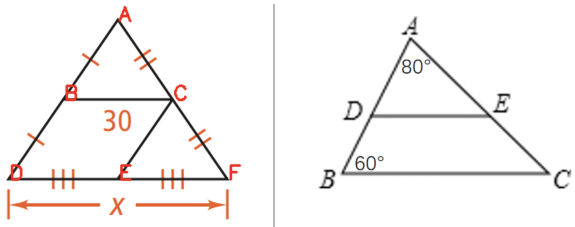
Table 6: Comparison with baseline methods on PGPS9K under Text-centric and Vision-centric evaluation settings.

Model	Text-centric	Vision	Avg
Qwen 2.5 7B Base Model	55.01	45.76	51.10
Qwen 2.5 7B MMICL	55.92	48.15	51.26
Qwen 2.5 7B PAPO	52.73	45.72	48.52
Qwen 2.5 7B KL (Ours)	**57.00**	**48.79**	**52.07**
Qwen 2.5 3B Base Model	35.68	28.66	31.47
Qwen 2.5 3B MMICL	43.80	36.21	39.24
Qwen 2.5 3B PAPO	47.87	40.57	43.49
Qwen 2.5 3B KL (Ours)	**48.19**	**41.58**	**44.23**

Table 7: Results on MathVerse under Text-centric and Vision-centric evaluation settings.

Model	MATH500	MMSTAR
Qwen 2.5 7B Base Model	65.80	60.47
Qwen 2.5 7B PAPO	59.20	**63.00**
Qwen 2.5 7B MMICL	**66.40**	61.00
Qwen 2.5 7B KL+Curriculum (Ours)	62.20	61.87
Qwen 2.5 3B Base Model	54.60	50.13
Qwen 2.5 3B PAPO	61.00	54.40
Qwen 2.5 3B MMICL	**64.60**	49.93
Qwen 2.5 3B KL+Curriculum (Ours)	59.40	**54.80**

Table 8: Results on general benchmarks MATH500 and MMSTAR.



PGPS9K
C is on line AF, E is on line DF, B is on line AD, BC = 30, BD = AB, AC = CF, EF = DE, DF = x. Find the value of x

MathVerse
In triangle ABC, it is known that angle A = 80.0, angle B = 60.0, point D is on AB and point E is on AC, DE parallel BC, then the size of angle CED is ()

Figure 4: Annotation style mismatch between PGPS9K and MathVerse. PGPS9K diagrams explicitly mark key geometric relations—parallelism and equality of segments/angles—whereas MathVerse omits such markings. Models trained on PGPS9K may over-rely on these visual tags and struggle to infer relations on MathVerse, weakening out-of-distribution generalization.

sults are shown in Table 6. Across both 3B and 7B model scales, our KL+Curriculum strategy outperforms generalized base models, MMICL-style in-context learning, and PAPO.

We further evaluate the same set of models on MathVerse, which tests multimodal mathematical reasoning beyond the training distribution. Results in Table 7 show that our method consistently achieves the best average performance across both text and vision settings at different model scales, demonstrating that the gains are not limited to PGPS9K.

Finally, we evaluate the same set of models on two widely used general benchmarks, MATH500 (Lightman et al., 2023) and MMSTAR (Chen et al., 2024). As shown in Table 8, training on PGPS9K with different optimization strategies does not degrade performance on general QA and VQA benchmarks. In particular, our KL-regularized curriculum training achieves performance comparable to existing baselines across both benchmarks, indicating that geometry-focused training on PGPS9K does not adversely affect the model's general reasoning and perception capabilities.

H Annotation Difference in Two Dataset

One key reason models trained on PGPS9K sometimes underperform on MathVerse is a mismatch in *annotation style*. As shown in Figure 4, PGPS9K explicitly marks geometric relations on the diagram—most notably ① segment parallelism and ② equivalence relations between segments and angles(e.g., equal-length segments and equal/corresponding/alternate angles). By contrast, MathVerse does *not* provide these markings. In several MathVerse settings, the model must infer these relations directly from the geometry without explicit visual tags, so a model trained on PGPS9K's fully annotated figures can overfit to those cues and exhibit weaker out-of-distribution generalization on MathVerse.

I Artifacts License

Our training codes primarily build upon the open-source training framework verl (Sheng et al., 2025), which is licensed under the Apache-2.0 License.

All source code developed for this work will be released under the Apache-2.0 License, which permits both research and commercial use, along

with modifications and distribution.

The two datasets used in this work, MathVerse (Zhang et al., 2024) and PGPS9K (Zhang et al., 2023) are licensed under MIT License, which allows for free use, modification, and distribution.

The Qwen2.5-VL series models (Bai et al., 2025) are released under the Apache-2.0 License, which permits both research and commercial use, along with modifications and distribution.