Supplementary Material for Understanding the Role of Hallucination in Reinforcement Post-Training of Multimodal Reasoning Models

Gengwei Zhang[1], Jie Peng[2], Zhen Tan[3], Mufan Qiu[1], Hossein Nourkhiz Mahjoub[4], Vaishnav Tadiparthi[4], Kwonjoon Lee[4], Yanyong Zhang[2], Tianlong Chen[1*]
[1] University of North Carolina at Chapel Hill [2] University of Science and Technology of China
[3] Arizona State University [4] Honda Research Institute, USA
*Correspondence to: Tianlong Chen <tianlong@cs.unc.edu>.

A. Implementation Details.

A.1. Evaluation Benchmarks

In our study, we evaluate model performance on four widely used visual mathematical reasoning benchmarks: **MathVision** [3], **MathVerse** [5], **MathVista** [1], and **We-Math** [2]. MathVision consists of 3,040 challenging mathematical problems across 12 grade levels, 16 subjects, and 5 difficulty tiers. MathVerse provides 3,940 well-defined problem–diagram pairs together with fine-grained categorizations of problem types. MathVista includes 1,000 samples covering a diverse range of visual mathematical reasoning tasks. WeMath contains 1,740 carefully curated problems spanning various knowledge granularities.

For evaluation across all benchmarks, we use the **Qwen2.5-32B-Instruct** model as an automated judge to determine the correctness of generated answers, with the ground-truth answer supplied to the judge for comparison.

A.2. Hallucination-as-Cue Implementation Details

We introduce three corruption strategies designed to encourage the model to reason under hallucination-inducing conditions: *Blank Image Replacement (BI)*, *Random Image Replacement (RI)*, and *Textual Information Removal (TR)*.

For *BI*, we replace the original image with a blank image of the same size, filled entirely with zeros.

For *RI*, we first remove the image corresponding to the problem from the dataset and randomly sample another image as input. As a result, both the content of the sampled image are typically unrelated to the problem text.

For *TR*, we design a rule-based parser to extract all condition statements and target information from the textual problem description and then remove them. However, due to the complexity of certain problems, the parser may fail. In such cases, we remove the entire problem description and keep only the system prompt that instructs the model to reason step-by-step and provide a final answer.

To isolate the individual effect of each corruption type, we apply only one corruption in either training or inference.

B. Fine-grained Analysis Details

To have a deeper understanding of the performance degradation observed across different evaluation datasets and to characterize the properties of the evaluation data, we conduct a fine-grained analysis based on question-type annotations from the MathVerse benchmark [5]. Specifically, we examine how RL-based post-training and modality-specific input corruptions affect model performance across four representative question categories: *Text-Dominant*, *Text-Lite*, *Vision-Intensive*, and *Vision-Dominant*. The definitions of these categories are as follows.

Text-Dominant. This variant retains as much textual information as possible, including descriptive content, implicit properties, and essential conditions. As a result, the text serves as the primary source of information required for problem solving.

Text-Lite. Compared to the *Text-Dominant* variant, this version removes most descriptive details, requiring the model to rely more on the accompanying diagram to infer basic information.

Vision-Intensive. Building upon the *Text-Lite* variant, this version further removes all implicit properties from the textual description. Consequently, the model must visually interpret mathematical relationships directly from the diagram.

Vision-Dominant. In this variant, essential problem conditions are explicitly annotated within the diagram itself. A model capable of perfectly interpreting visual information would perform equivalently to how it performs on the *Text-Lite* variant.

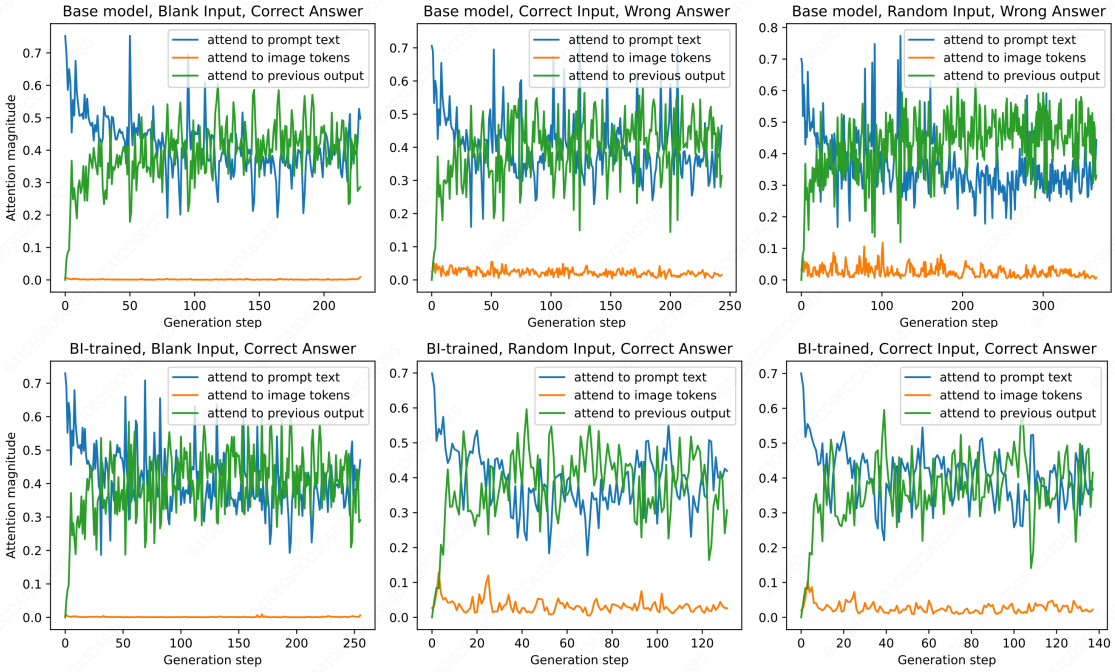

Figure 1. Attention magnitudes across generation steps for different types of tokens.

Table 1. MM-UPT Training Results.

Model	Training Dataset	MathVision (%)	MathVerse (%)	MathVista (%)	WeMath (%)	AVG (%)
Qwen2.5-VL-7B	–	27.70	45.20	67.00	63.68	50.89
GRPO	Geometry3K	28.13	47.56	70.00	68.39	**53.52**
GRPO	MMR1-V0	30.03	50.10	70.80	70.29	**55.31**
MM-UPT	Geometry3K	27.33	42.46	68.50	66.61	51.23
MM-UPT	MMR1-V0	26.15	44.87	72.90	68.74	53.17
MM-UPT-BI	Geometry3K	26.48	47.72	69.50	64.25	<u>51.99</u>
MM-UPT-BI	MMR1-V0	26.55	46.52	72.00	69.08	<u>53.54</u>

C. Additional Analysis

C.1. Discussion of Findings

1. Why corrupted training still improves performance? Using the case study from Figure 1 in our main paper as an example, from the perspective of reasoning, the small portion of positive hallucinated trajectories do exhibit correct textual reasoning. Rewarding such trajectories encourages the model to learn effective reasoning behaviors, while negative hallucinated trajectories are discouraged, preventing excessive hallucination.

2. Why do larger models show larger improvements? Larger models have greater capacity to generate and learn from hallucinated positive trajectories. Concretely, the 7B model has an initial hallucinated accuracy of 9.7% on Geometry3K-BI, which increases to 14.1% after training, while the 3B model improves from 7.6% to 10.4%.

3. Why in some cases surpass standard training? The underlying mechanism is complex and multifaceted, and this is one of the key points that we aim to draw the community's attention to for further investigation. Here we provide one possible explanation. Although standard training receives more effective reward signals, this does not necessarily lead to better optimization outcomes. A key reason is that standard training admits a much broader set of positive trajectories, including many low-quality but reward-satisfying ones (e.g., descriptive rather than reasoning), which can dilute optimization pressure. In contrast, corrupted training substantially restricts the space of viable positive trajectories. When essential information is missing, only a small subset of trajectories can still achieve positive rewards.

C.2. Attention Visualization

Fig. 1 visualizes the attention distributions over different token types. The analysis is based on the case study presented in Figure 1 of the main paper. The results show that, after training, the model still adaptively attends to visual tokens rather than ignoring them. Under corrupted inputs, the

Table 2. Reward Ablation Study for Qwen2.5-VL-3B + GRPO-BI.

Model	Reward	MathVision (%)	MathVerse (%)	MathVista (%)	WeMath (%)	AVG (%)
Qwen2.5-VL-3B	–	18.19	34.82	51.40	54.48	39.72
Qwen2.5-VL-3B + GRPO-BI	format+accuracy	20.95	35.10	56.40	56.55	42.25
Qwen2.5-VL-3B + GRPO-BI	accuracy	20.82	35.56	58.00	57.64	**43.01**
Qwen2.5-VL-3B + GRPO-BI	format	19.74	35.46	55.40	53.28	40.97

Table 3. Effect of Rollout Size. We use Qwen2.5-VL-3B as the base model for both GRPO and GRPO-BI training.

Model	Rollout Size	MathVision (%)	MathVerse (%)	MathVista (%)	WeMath (%)	AVG (%)
GRPO	5	28.13	47.56	70.00	68.39	53.52
GRPO	10	27.27	51.98	72.40	68.10	54.94
GRPO	15	28.06	50.81	71.30	69.83	55.00
GRPO-BI	5	28.39	48.86	68.50	66.84	53.15
GRPO-BI	10	28.22	49.31	69.90	66.44	53.47
GRPO-BI	15	27.93	49.47	70.40	67.59	53.85

model compensates for missing information through hallucination, while when sufficient information is available, it relies on the provided inputs for reasoning.

D. Additional Results

D.1. Hallucination-as-Cue in Unsupervised Training

Recently, MM-UPT [4] demonstrated that majority-voting over sampled model outputs can serve as effective pseudo-labels, enabling unsupervised RL-based multimodal reasoning post-training. We incorporate our Hallucination-as-Cue framework into this unsupervised training paradigm to investigate whether hallucination-dominant trajectories can also enhance the reasoning capability of multimodal models when ground-truth supervision is imperfect. The corresponding results are shown in Tab. 1.

Notably, we observe that under the *BI* corruption setting, RL-based training achieves performance comparable to that of the non-corrupted unsupervised training baseline.

D.2. Effect of Formatting Reward

In most RL-based post-training pipelines for reasoning models, a *formatting reward* is applied alongside the accuracy reward. This reward encourages the model to structure its reasoning output in a predefined format, such as enclosing intermediate reasoning within a "`<thinking></thinking>`" block. To decouple the impact of this formatting reward with accuracy reward, we conduct an ablation study using the Qwen2.5-VL-3B + GRPO-BI model. The results are presented in Tab. 2.

D.3. Effect of Rollout Size

In GRPO-based post-training, a larger rollout (group) size provides a more reliable estimate of the expected reward by reducing the variance of sampled trajectories. To assess its influence, we experiment with rollout sizes of 5, 10, and 15 for both the standard GRPO model and our GRPO-BI variant. The results are summarized in Tab. 3.

Overall, we observe that increasing the rollout size consistently improves the performance of both GRPO and GRPO-BI models. However, the relative performance gains for GRPO-BI are smaller compared to standard GRPO training without corruption, suggesting that hallucination-inducing conditions reduce the marginal benefit of additional rollouts.

References

[1] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1

[2] Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070, 2025. 1

[3] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 1

[4] Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. First sft, second rl, third upt: Continual improving multi-modal llm reasoning via unsupervised post-training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[5] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu

Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 1