Mitigating Information Forgetting via Entropy-Driven Progressive Retrospection for Multimodal Long Reasoning

Supplementary Material

A. More Experiment Details

A.1. Detailed Introduction of Base Models

In this section, we provide a more detailed overview of the three SOTA Multimodal Large Language Models (MLLMs) used as base models in our experiments. These models are built upon diverse and distinct training diagrams, providing a robust evaluation for the generalizability of our EDPR framework.

VLAA-Thinker-Qwen2.5-VL [3] uses *a pure RL approach, entirely bypassing SFT stage*. The core idea of VLAA-Thinker is that SFT may induce the model to imitate pseudo-reasoning paths from expert models, including hesitations or flawed logic, which can limit the exploratory potential of the subsequent RL stage. The VLAA-Thinker is trained directly on base Qwen2.5-VL using GRPO. Its training is guided by a hybrid reward module that combines rule-based rewards (e.g., numerical matching, bounding box IoU) and an open-ended reward signal from a powerful reward model. This pure RL paradigm aims to drive the model towards discovering authentic and adaptive reasoning behaviors from scratch.

OpenVLThinker [8] employs *an iterative SFT-RL curriculum*. This framework is designed to solve the twin challenges of imprecise visual grounding in pure SFT and an overly large search space in pure RL. The core idea is a self-improving loop: an initial, lightweight SFT stage uses CoTs generated by a powerful text-based R1 reasoning model QwQ-32B to bootstrap the model's reasoning behavior and narrow the exploratory search space for RL. Subsequently, a curriculum-based RL stage (also using GRPO) fine-tunes the model on data of increasing difficulty. The optimized model from the RL stage is then used to generate higher-quality data for the next iteration of SFT. This iterative SFT-RL curriculum allows the model to progressively refine its reasoning capabilities with high data efficiency.

MM-EUREKA-Qwen [20] is fine-tuned with *a stable, rule-based RL methodology using GRPO*, with a strong focus on high-quality and human-verified in-domain data. Its primary training dataset MMK12, consists of K-12 level mathematics problems. To ensure training stability, which is a common challenge in RL for LLMs, MM-EUREKA-Qwen employs several key techniques. It uses an online filtering strategy to dynamically remove samples that provide no learning signal (i.e., those that are either always correct or always incorrect). The reward system is sparse and rule-based, focusing on two clear objectives: accuracy reward (a binary score for the final answer) and format reward (a score for adhering to the specified output format). This focus on stability and high-quality, domain-specific data aims to improve deep and reliable reasoning abilities in a specific vertical.

A.2. Generation Settings

To ensure fair and reproducible comparisons, we employ consistent generation parameters for all experiments. We use widely-used temperature of 0.3 and top-p value of 0.95 to ensure output stability and achieve controllable diversity. In addition, we employ a common repetition penalty of 1.05 to reduce repeated generation.

Moreover, for VLAA-Thinker-Qwen2.5-VL and MM-Eureka-Qwen, we follow their officially recommended system prompts for their optimal reasoning capabilities and formatted outputs.

VLAA-Thinker-Qwen2.5-VL System Prompt

You are VL-Thinking, a helpful assistant with excellent reasoning ability. A user asks you a question, and you should try to solve it. You should first think about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within $< think >< /think >$ and $< answer >< /answer >$ tags, respectively, i.e., $< think >$ reasoning process here $< /think >< answer >$ answer here $< /answer >$

MM-Eureka-Qwen System Prompt

Solve the question. The user asks a question, and you solves it. You first thinks about the reasoning process in the mind and then provides the user with the answer. The answer is in latex format and wrapped in $...$. The final answer must be wrapped using the boxed{} command. The reasoning process and answer are enclosed within $< think >< /think >$ and $< answer >< /answer >$ tags, respectively, i.e., $< think >$ Since $1 + 1 = 2$, so the answer is 2. $< /think >< answer >$ The answer is boxed{2} $< /answer >$, which means assistant's output should start with $< think >$ and end with $< /answer >$.

Table 4. Ablation study on the effect of triggering shreshold and throttling distance, conducted on MathVista with MM-EUREKA-Qwen as base model.

Threshold Factor	Accuracy	Throttling Distance	Accuracy
distance=100		threshold=150	
100	72.80	50	73.50
150	**73.60**	**100**	**73.60**
200	73.20	150	73.40

B. More Ablation Studies

In this section, we provide additional ablation studies to further validate the effect of other hyperparameters in our EDPR framework.

In addition to the attention operation, the precise and efficient intervention triggering mechanism also has effect on our EDPR. To determine their optimal values and understand their impact, we conduct ablation studies on MathVista benchmark with MM-EUREKA-Qwen-7B as the base model. The results are presented in Tab.4.

Effect of Triggering Threshold. The results on the left of Tab.4 show that the trigger sensitivity has a notable impact on performance. The accuracy exhibits a distinct optimal point. A lower threshold (e.g., 100) results in lower accuracy (72.80%), likely because the detector is overly sensitive and triggers interventions on minor, less critical entropy fluctuations, introducing unnecessary computational overhead and potentially distracting the model. Conversely, a higher threshold (e.g., 200) also leads to a performance drop (73.20%), as it risks becoming too insensitive and failing to intervene at genuine moments of high uncertainty. A Triggering Threshold of 150 achieves the best performance (73.60%), striking an effective balance between sensitivity to critical reasoning junctures and robustness against noise.

Effect of Throttling Distance. The results on the right of Tab.4 explore the effect of distance between interventions. The model demonstrates considerable robustness to this parameter, with performance remaining highly stable across the tested range (73.40% to 73.60%). This suggests that while a throttling mechanism is crucial to prevent redundant interventions, it is mainly used to provides a reasonable window for previous processing, less critical than the trigger's sensitivity.

Robustness Under Sub-Optimal Settings. Our comprehensive ablation studies in Fig.5 and Tab.3,4 show that our EDPR surpass the base model (MM-EUREKA-Qwen-7B: 72.80%) even under most sub-optimal hyperparameter settings. This result indicates that the core mechanism of progressive information retrospection in our EDPR is fundamentally effective. Even when the intervention is applied at sub-optimal strengths or in less ideal layers, the context-aware revisitation for crucial multimodal evidence at high-entropy moments provides a consistent and positive impact. This robustness confirms that mitigating information forgetting through precise intervention is a effective and reliable strategy for enhancing multimodal long reasoning.