First SFT, Second RL, Third UPT: Continual Improving Multi-Modal LLM Reasoning via Unsupervised Post-Training

Lai Wei, *Yuting Li, Chen Wang, Yue Wang, Linghe Kong,*
Weiran Huang, Lichao Sun

NeurIPS 2025

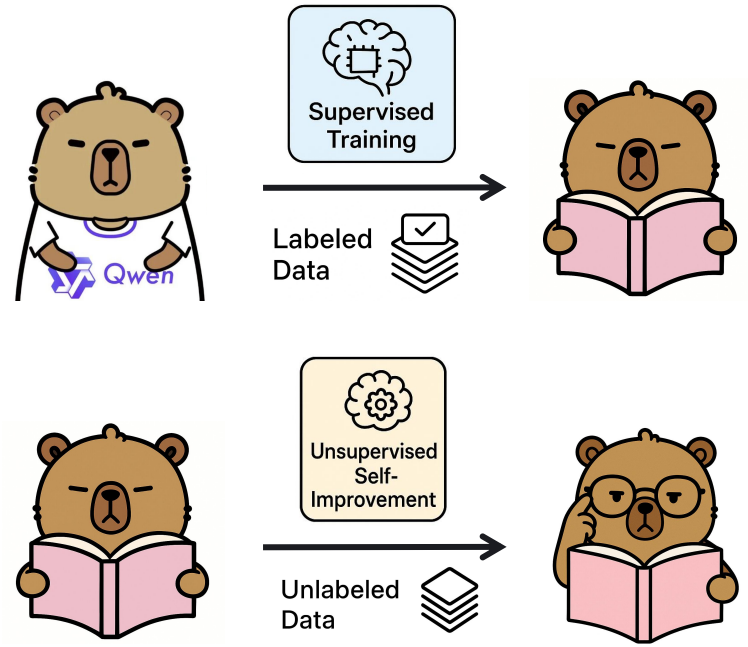
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

MIFA LAB
机器智能基础与应用实验室

Motivation

Reinforcement learning requires expensive and manually annotated multi-modal data–**an ultimately unsustainable resource**.

To overcome this data-dependency, a paradigm shift is required towards a **third stage** of post-training beyond SFT and RL, dedicated to the continual self-improvement of MLLMs through synthetic and unlabeled data. We formalize this third-stage paradigm as **Unsupervised Post-Training**.



Previous Works

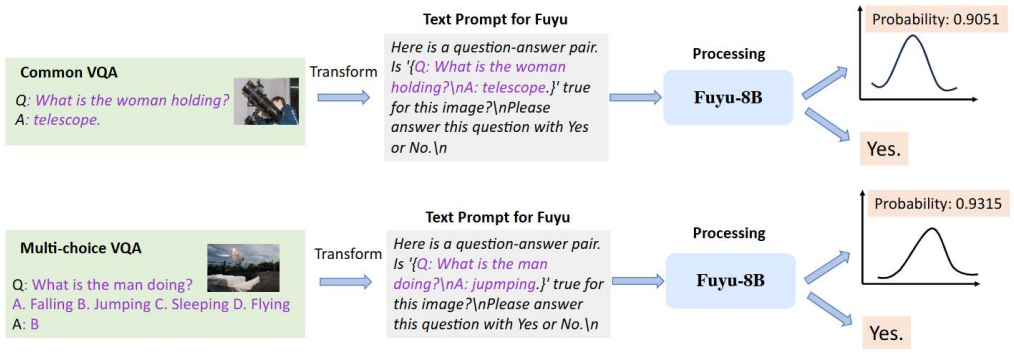


Fig. 5: The illustration of proposed Fuyu-driven data filtering framework. The outputs of the framework compose a probability and a direct answer.

Zhao, Henry Hengyuan, Pan Zhou, and Mike Zheng Shou. "Genixer: Empowering multimodal large language model as a powerful data generator." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

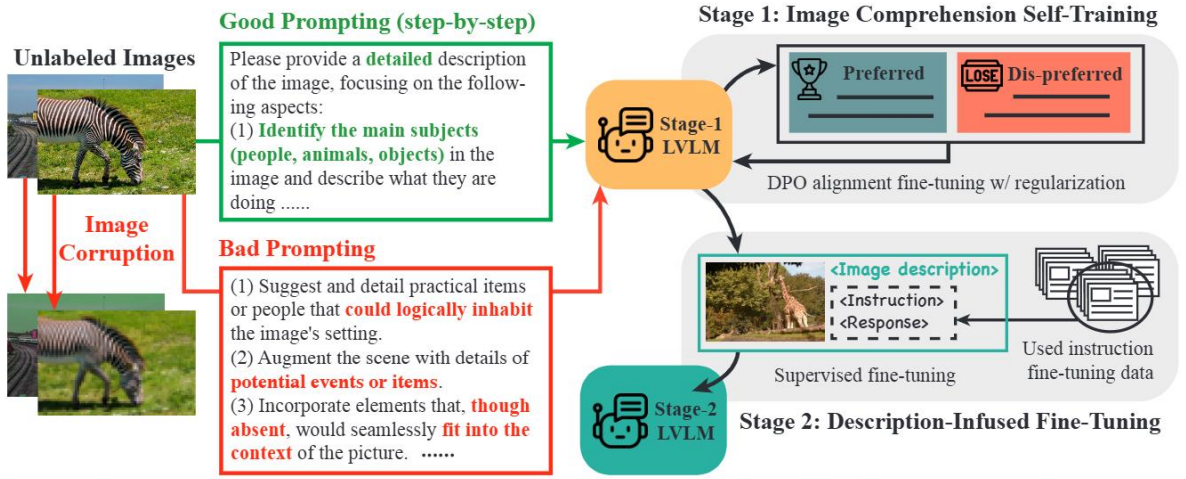


Figure 2: Framework overview of STIC, a two-stage self-training algorithm focusing on the image comprehension capability of the LVLMs. In Stage 1, the base LVLM self-constructs its preference dataset for image description using well-designed prompts, poorly-designed prompts, and distorted images. In Stage 2, a small portion of the previously used SFT data is recycled and infused with model-generated image descriptions to further fine-tune the base LVLM.

Deng, Yihe, et al. "Enhancing large vision language models with self-training on image comprehension." Advances in Neural Information Processing Systems 37 (2024): 131369-131397.

Training Method

MM-UPT introduces a self-rewarding mechanism using majority voting as pseudo-labels based on the online reinforcement learning. This strategy is also adopted in some concurrent works, such as TTRL and SRT.

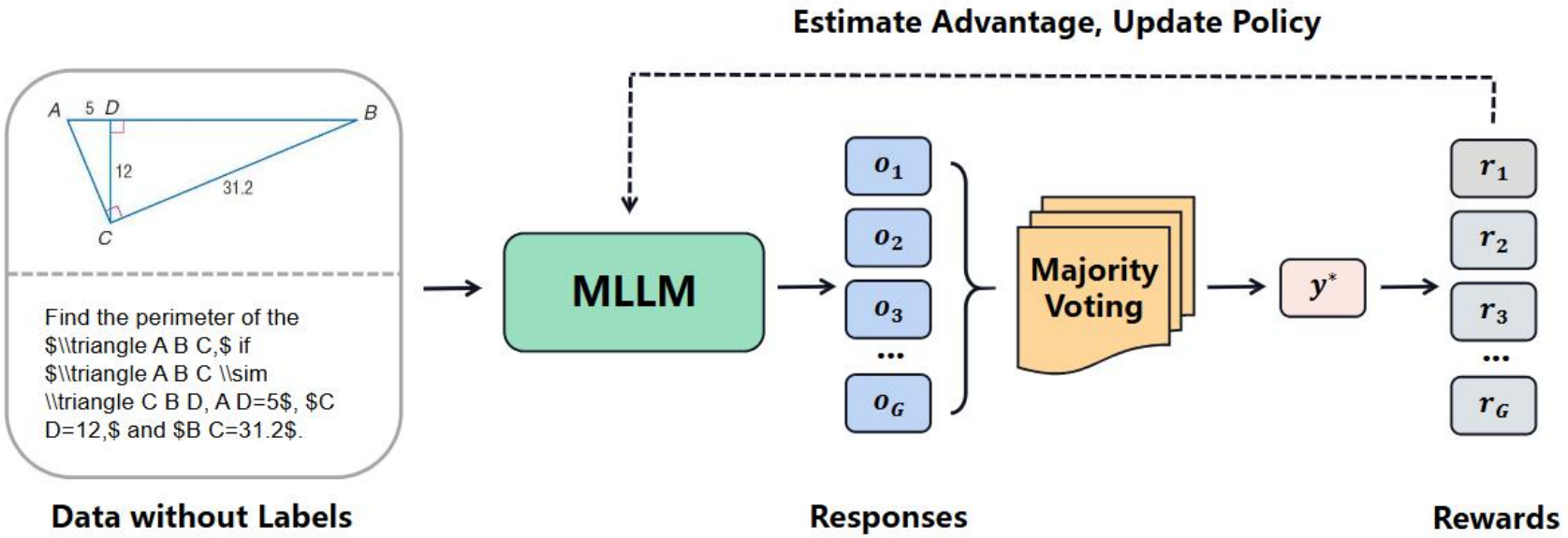


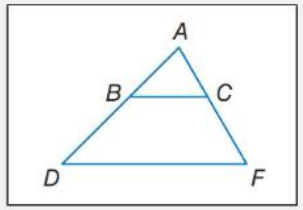
Figure 1: Overview of the MM-UPT framework. Given an unlabeled multi-modal input, the MLLM samples multiple responses, and uses majority voting to determine the pseudo-label. The MLLM is then updated via GRPO, enabling self-improvement without external supervision.

Data Synthesis

In-Context Synthesizing.
To synthesize new samples, we provide the model with the full triplet and instruct it to generate a new question that is semantically distinct from the original but relevant to the same image. This strategy helps generate task-relevant and meaningful variations of the original question, as well as ensure the quality of synthetic questions.

Direct Synthesizing.
Here, the model receives only the image and is prompted to freely create a new question without any reference to the original question. This open-ended formulation encourages the model to generate a wider range of diverse and novel questions based solely on the visual input, rather than being constrained by the original task.

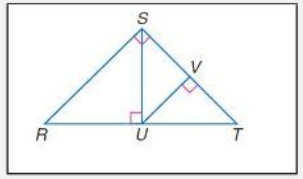
Demo: Examples of synthetic data using different strategies.



Original Question:
$BC \| DF$. $AB = x + 5$, $BD = 12$, $AC = 3x + 1$, and $CF = 15$. Find x.

In-Context Synthetic Question:
Given that $BC \| DF$, $AB = 2x - 3$, $BD = 18$, $AC = x + 7$, and $CF = 24$. Find the value of x.

Directly Synthetic Question:
In the given triangle $\triangle ADF$, point B lies on AD and point C lies on AF. If $BC \parallel DF$, what is the ratio of the area of $\triangle ABC$ to the area of $\triangle ADF$?



Original Question:
If $\angle RST$ is a right angle, $SU \perp RT$, $UV \perp ST$, and $\angle RTS = 47°$, find $\angle RSU$

In-Context Synthetic Question:
If $\angle RST$ is a right angle, $SU \perp RT$, $UV \perp ST$, and $\angle RTS = 47°$, find $\angle VST$.

Directly Synthetic Question:
In the given triangle $\triangle RST$, point U lies on RT such that SU is perpendicular to RT. Point V lies on ST such that UV is perpendicular to ST. If $RU = 12$ units, $UT = 16$ units, and $SV = 9$ units, find the length of VT.

Main Experiments

Table 1: Main results on four multi-modal mathematical reasoning benchmarks. We report accuracy (%) for each method on MathVision, MathVerse, MathVista, and We-Math. All methods are conducted on the Qwen2.5-VL-7B backbone. MM-UPT outperforms other baseline methods, and is even competitive with supervised methods.

Model and Methods	Unsupervised?	Training Data	MathVision	MathVerse	MathVista	We-Math	Avg
Qwen2.5-VL-7B	-	-	24.87	43.83	66.30	62.87	49.47
+ GRPO [33]	✗	Geometry3K	28.32	46.40	69.30	68.85	53.22
+ GRPO [33]	✗	GeoQA	26.15	46.28	67.50	66.65	51.65
+ GRPO [33]	✗	MMR1	29.01	45.03	71.40	67.24	53.17
+ SFT [38]	✗	Geometry3K	25.92	43.73	67.90	64.94	50.63
+ SFT [38]	✗	GeoQA	25.72	44.70	67.40	65.10	50.73
+ SFT [38]	✗	MMR1	26.45	43.53	63.30	64.20	49.37
+ SRLM [48]	✓	Geometry3K	26.94	44.54	66.90	66.32	51.18
+ SRLM [48]	✓	GeoQA	25.16	44.62	66.30	65.00	50.27
+ SRLM [48]	✓	MMR1	25.33	45.08	67.00	64.66	50.52
+ LMSI [13]	✓	Geometry3K	25.10	43.96	65.50	64.43	49.75
+ LMSI [13]	✓	GeoQA	25.49	43.50	66.60	63.51	49.78
+ LMSI [13]	✓	MMR1	24.83	43.76	64.90	66.38	49.97
+ Genixer [61]	✓	Geometry3K	26.02	43.15	65.50	62.18	49.22
+ Genixer [61]	✓	GeoQA	25.30	44.11	66.80	64.25	50.12
+ Genixer [61]	✓	MMR1	23.68	43.30	65.50	64.66	49.29
+ STIC [5]	✓	Geometry3K	25.39	42.92	65.20	62.99	49.13
+ STIC [5]	✓	GeoQA	23.49	42.87	64.30	63.62	48.57
+ STIC [5]	✓	MMR1	23.78	42.72	66.10	63.74	49.09
+ MM-UPT	✓	Geometry3K	27.33	42.46	68.50	66.61	**51.23**
+ MM-UPT	✓	GeoQA	27.07	43.68	68.90	68.22	**51.97**
+ MM-UPT	✓	MMR1	26.15	44.87	72.90	68.74	**53.17**

Table 2: Performance comparison of MM-UPT using different synthetic data generation strategies. Both "In-Context Synthesizing" and "Direct Synthesizing" approaches yield significant improvements over the base model and perform competitively with the "Original Questions" on average, demonstrating the effectiveness of synthetic data for unsupervised self-improvement.

Model and Methods	MathVision	MathVerse	MathVista	We-Math	Avg
Qwen2.5-VL-7B	24.87	43.83	66.30	62.87	49.47
w/ Original Questions	27.33	42.46	68.50	66.61	51.23 (**3.6%**↑)
w/ In-Context Synthesizing	26.71	41.24	68.30	67.76	51.00 (**3.1%**↑)
w/ Direct Synthesizing	26.88	43.53	69.90	68.97	52.32 (**5.8%**↑)

Our experiments are designed to explore **two key scenarios**:

(1) using human-created questions without ground-truth labels

(2) employing synthetic questions generated by the model itself, inherently lacking ground-truth labels

Table 1: Unsupervised post-training on unlabeled datasets significantly improves multimodal reasoning ability of Qwen2.5-VL-7B：

- MathVista：66.3% → 72.9%; We-Math：62.9% → 68.7%
- Outperforms all existing unsupervised methods.
- Approaches the performance of supervised GRPO.

Table 2: Training with self-synthetic data also delivers significant gains. It demonstrates a scalable self-improvement paradigm.

Further Experiments

Table 3: Ablation study using different models besides Qwen2.5-VL-7B. We conduct this experiment on Geometry3K [25] dataset without labels.

Models	MathVision	MathVerse	MathVista	We-Math	Avg
Qwen2.5-VL-7B	24.87	43.83	66.30	62.87	49.47
Qwen2.5-VL-7B + MM-UPT	27.33	42.46	68.50	66.61	51.23 (**3.6%**↑)
MM-Eureka-7B	28.06	50.46	69.40	64.48	53.10
MM-Eureka-7B + MM-UPT	28.95	50.63	69.10	66.44	53.78 (**1.3%**↑)
ThinkLite-VL-7B	26.94	46.58	69.00	67.99	52.63
ThinkLite-VL-7B + MM-UPT	26.91	47.26	74.70	67.41	54.07 (**2.8%**↑)
Qwen2.5-VL-3B	19.47	33.58	56.30	50.63	39.00
Qwen2.5-VL-3B + MM-UPT	22.17	32.39	57.10	55.22	41.72 (**7.4%**↑)

Table 6: Performance on non-mathematical VQA benchmarks. We evaluate Qwen2.5-VL-7B before and after applying MM-UPT on the MMR1 dataset. Scores are reported as accuracy.

Models	ChartQA	IconQA
Qwen2.5-VL-7B	71.96	54.20
Qwen2.5-VL-7B + MM-UPT	77.48 (**7.7%**↑)	56.55 (**4.3%**↑)

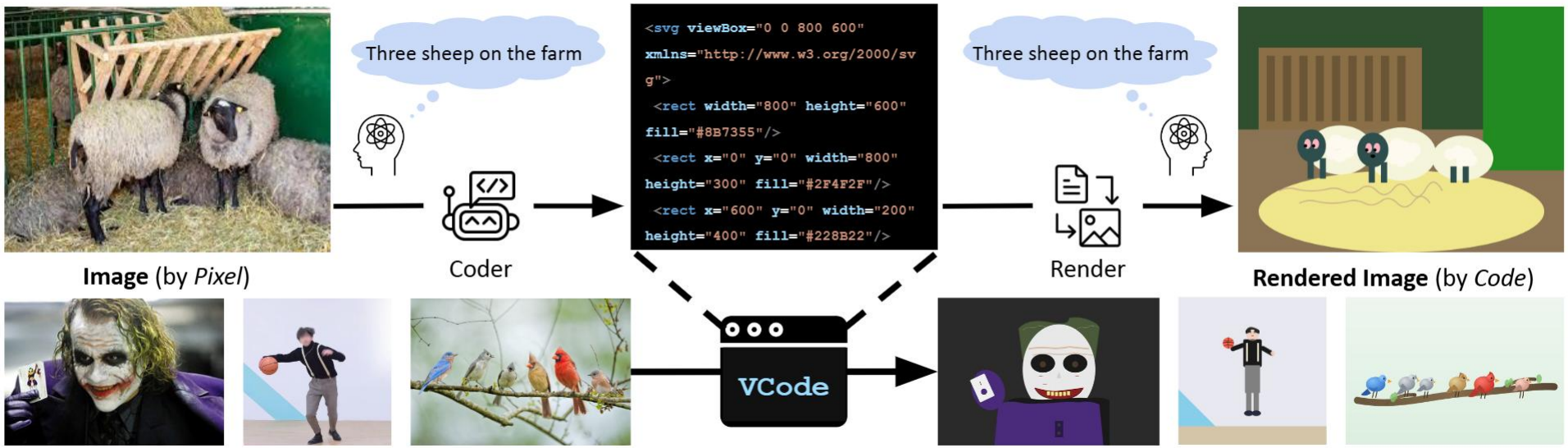
Table 3: MM-UPT can be easily applied to various multi-modal models to enable consistent self-improvement. These results also show that MM-UPT is compatible with supervised GRPO.

Table 6: To investigate whether the method suffers from a negative impact on broader generalization, we extend our evaluation to two non-mathematical visual question answering benchmarks.

Future Works

1. Combining MM-UPT with more self-rewarding algorithms (such as LLM-as-a-Judge) and data synthesis methods (such as Text2SVG) will be a promising direction.

2. Meanwhile, exploring the scaling laws of unsupervised post-training also represents an interesting avenue for further research.



Thank you

Paper

Github