OpenMMReasoner: Pushing the Frontiers for Multimodal Reasoning with an Open and General Recipe

Supplementary Material

Training Component	SFT	RL
optimizer	AdamW	AdamW
scheduler	cosine	constant
learning rate	5e-5	1e-6
weight decay	0.0	0.1
Training Steps	4300	1232
Warmup Steps	430	25
Max Length	61440	32792
Dynamic Bsz	True	True
Remove padding	True	True
Liger Kernel	True	False

Table 1. Detail parameters for SFT and RL training.

1. Implementation Details

1.1. Training Details

We describe the training procedures for our best-performing models in both the SFT and RL stages.

SFT. To maximize training throughput and reduce memory consumption, we apply online stream packing with an iterable dataset, removing padding tokens to avoid unnecessary computation. We use a packing length of 61,440 tokens. For each batch, we dynamically compute the input token lengths of incoming samples and fill the batch buffer until it is fully packed. Because packing is performed online, the total number of epochs cannot be predetermined; instead, we train each model until convergence. The full set of experimental hyperparameters is provided in Tab. 1.

RL. For RL training, we use a global batch size of 128, which strikes a balance between on-policy stability and training speed. Log probabilities are computed using a dynamic batch-size implementation without padding to further reduce memory usage. During generation, we set the maximum number of new tokens to 28,696 and cap the prompt length at 4,096 tokens, resulting a max length at 32792. We use a temperature of 1.0 and keep this configuration fixed across all experiments. Due to the high computational cost of RL training, we run training until the reward saturates. The detailed hyperparameters are listed in Tab. 1.

1.2. Evaluation Details

We now describe our evaluation setup for multimodal reasoning benchmarks. The SFT and RL models share the same evaluation configuration. We use the system prompt shown in Tab. 6 to ensure the model outputs both the reasoning trace and the final answer in an extractable format. The extracted answer is then validated using a two-stage process: a rule-based validator followed by an LLM-as-judge validator. We first apply the rule-based validator to minimize evaluation cost; if the answer cannot be verified, we fall back to the LLM-as-judge, using the prompt provided in Tab. 7. For all evaluations, we use a temperature of 0.0 for reproducibility and set the maximum generation length to 49K tokens, except for AIME, where we use a temperature of 1.0. We employ vLLM [17] as the serving engine to accelerate inference.

2. Additional Evaluation Details

To assess the generalization capability of our method beyond mathematical reasoning, we conduct comprehensive evaluations on a diverse set of cross-domain multimodal benchmarks using lmms-eval.

General Multimodal Benchmarks. We first evaluate on benchmarks that assess broad multimodal understanding and reasoning capabilities. As shown in Tab. 4, OMR achieves 83.4% on AI2D (scientific diagrams) and 89.9% on MMBench, outperforming the Qwen2.5-VL-7B baseline by 5.5 and 2.7 points respectively. On RealWorldQA, which tests real-world spatial and physical reasoning, our method maintains competitive performance at 51.0%, demonstrating that the improvements on reasoning tasks do not come at the cost of general perception abilities.

Domain-Specific Reasoning Benchmarks. We further evaluate on benchmarks that require specialized reasoning across different domains. As presented in Tab. 5, OMR demonstrates substantial improvements on PhyX (physics reasoning), achieving 20.4% on the open-ended split and 55.9% on the multiple-choice split, compared to 14.5% and 28.4% from the baseline. On OlympiadBench, which contains competition-level problems, our method improves performance from 20.4% to 37.3%. Notably, these gains come without degradation on OCRBenchV2, where we maintain comparable performance at 48.0%.

These results demonstrate that the proposed SFT+RL training recipe generalizes effectively beyond the mathematical domain, yielding transferable improvements across diverse multimodal reasoning tasks including scientific

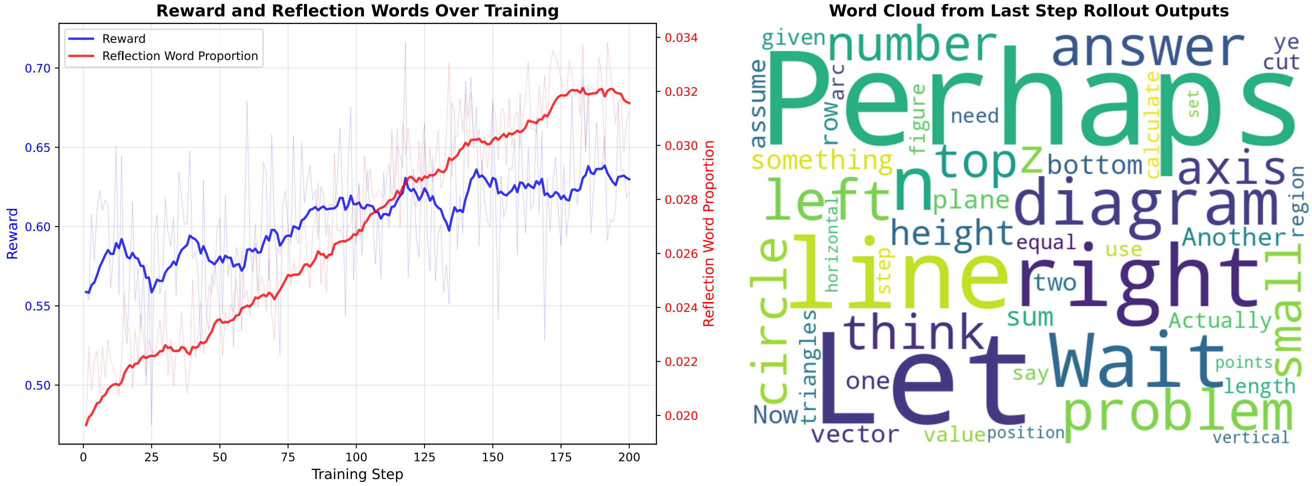


Figure 1. **Rollout Analysis over RL.** With the progress of RL training, the model response contains more reflection word ratio.

Table 2. **Additional evaluation results on Sampling Strategy.** Scaling up with more answer generation leads to better result.

Method	**Avg**	**MMMU** val	**MMMU_Pro** standard	**MMMU_Pro** vision	**MathVista** testmini	**MathVerse** test	**MathVision** testmini	**MathVision** test	**CharXiv** reas	**CharXiv** desc	**LogicVista** test	**WeMath** loose	**DynaMath** worst
Sampling Strategy													
$\times 1$ sampling	46.7	54.2	42.6	36.0	71.5	49.7	29.9	28.2	39.0	73.1	47.8	63.9	25.0
$\times 2$ sampling	47.5	54.4	39.1	34.6	72.9	54.4	30.9	30.8	37.5	73.6	49.6	69.3	23.0
$\times 4$ sampling	48.2	56.9	41.0	37.6	72.6	55.3	30.3	30.8	40.7	75.4	45.8	69.3	23.0
$\times 8$ sampling	49.2	55.6	40.6	37.7	73.7	57.1	35.2	34.6	39.5	74.1	48.9	69.1	24.0

diagram understanding, physics problem solving, and competition-level reasoning, while preserving core perception capabilities such as OCR.

3. Additional Result and Analysis

Sampling Scaling Results. We present the full evaluation results for different sampling strategies in Tab. 2. As shown in the table, scaling up the sampling strategy improves the average score from 46.7 to 49.2, demonstrating the effectiveness of increasing answer diversity along this axis.

Rollout Analysis. During RL training, we record all rollout logs and analyze the proportion of reflection-related words in the model's outputs. We observe that as the reward increases over training steps, the proportion of reflection words also rises. To illustrate this, we generate a word cloud from the final-step rollout outputs—after removing noise words—and find that reflection cues such as "let," "wait," and "think" appear frequently in the model's responses. This indicates that our RL recipe effectively encourages the model to engage in more explicit reasoning as its capabilities improve. The results are shown in Fig. 1.

Reward λ_{fmt} Ablation. To assess the impact of the λ_{fmt} parameter in RL, we conduct additional experiments using four values: 0.1, 0.3, 0.5, and 0.7, under the same experimental setup as the GRPO configuration. The evaluation results are presented in Tab. 3. We observe that a lower format-reward weight, specifically $\lambda_{fmt} = 0.1$, consistently yields the best performance. Consequently, we adopt $\lambda_{fmt} = 0.1$ in our final configuration.

4. Examples

Data examples. We present several examples in Tab. 8 and Tab. 9 from our reasoning dataset to demonstrate the high quality of the answer traces generated by our method. These cases illustrate how the model decomposes complex questions into structured reasoning steps and produces coherent, verifiable conclusions.

Qualitative Results. We present several qualitative examples in Tab. 10 and Tab. 11 to demonstrate the effectiveness and robustness of our model. These examples highlight the model's ability to accurately interpret complex multimodal inputs, generate coherent reasoning steps, and produce reliable answers across diverse scenarios, showcasing both its generalization capability and practical utility.

Table 3. **Reward ablation result.**

Method	Avg	MMMU	MMMU_Pro		MathVista	MathVerse	MathVision		CharXiv		LogicVista	WeMath	DynaMath
		val	standard	vision	testmini	test	testmini	test	reas	desc	test	loose	worst
Reward λ_{fmt} Settings													
$\lambda_{fmt} = 0.1$	51.1	54.6	42.8	39.4	77.1	58.3	33.9	37.1	43.1	73.8	51.1	70.2	32.1
$\lambda_{fmt} = 0.3$	47.0	51.3	36.7	33.9	73.9	55.8	29.6	34.1	36.5	75.9	43.3	64.3	29.1
$\lambda_{fmt} = 0.5$	45.0	48.3	34.3	32.4	74.2	53.1	26.0	27.9	39.8	75.6	41.1	58.1	29.3
$\lambda_{fmt} = 0.7$	48.8	53.4	37.8	35.8	74.8	56.8	32.9	35.8	39.4	75.7	46.9	66.9	29.1

Split	AI2D test	MMBench dev	RealworldQA test
Qwen2.5-VL-7B	77.9	87.2	53.3
OMR(ours)	**83.4**	**89.9**	51.0

Table 4. Evaluation comparison on general multimodal benchmarks evaluating cross-domain reasoning and robustness.

Split	Phyx oe	Phyx mc	OlympiadBench test	OCRBenchV2 test
Qwen2.5-VL-7B	14.5	28.4	20.4	48.3
OMR-RL	**20.4**	**55.9**	**37.3**	48.0

Table 5. Evaluation comparison on physics, competition-style, and OCR-focused multimodal reasoning benchmarks.

5. Limitation and Future Work

Our work primarily focuses on a single model family—Qwen2.5-VL-Instruct—and evaluates performance mainly within the image domain. While our approach demonstrates strong gains in multimodal reasoning, it does not extend to other modalities such as video or audio, limiting its applicability to a broader set of real-world tasks. Additionally, although we explore scaling strategies in both SFT and RL stages, we have not yet identified the upper bound of model performance under further scaling, leaving open the question of how far the current recipe can be pushed.

For future work, we aim to extend our methodology to a wider range of domains, including video, audio, and richer temporal-stream modalities. Another promising direction is to explore generation capabilities across multiple modalities simultaneously, enabling more coherent and context-aware multimodal reasoning and content synthesis. Broadening the model family used for evaluation and training will further help validate the generality and robustness of our findings across architectures and modality configurations.

System Prompt for model

```
You are a helpful assistant. When the user asks a question, your response must include two parts: first, the reasoning process
    ↪ enclosed in <think>...</think> tags, then the final answer enclosed in <answer>...</answer> tags. Please provide a
    ↪ clear, concise response within <answer> </answer> tags that directly addresses the question.
```

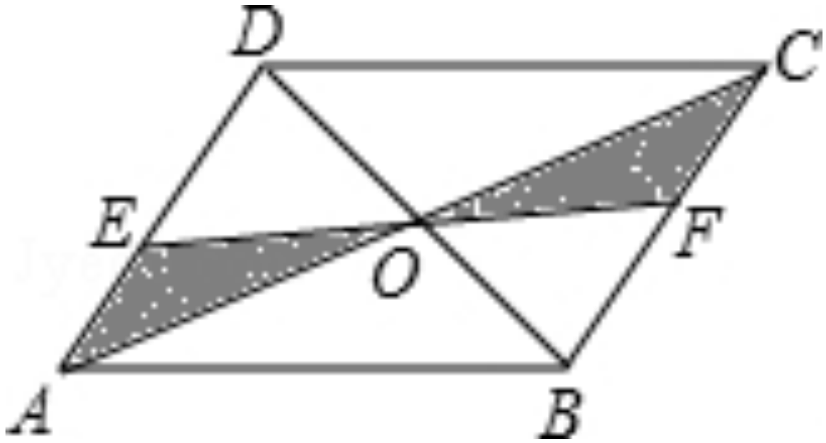
Table 6. The prompt that used in evaluation.

System Prompt for judge model

```
You are a strict evaluator assessing answer correctness. You must output 1 for fully correct answers and 0 for any other case.
    ↪ You will receive the question, the ground truth answer, and the model prediction.

# Input
Question:
```
{question}
```

Ground Truth Answer:
```
{answer}
```
Model Prediction:
```
{prediction}
```

# Evaluation Rules
- For multiple-choice questions: Score 1 if the predicted answer matches the ground truth answer, it can be directly in option
    ↪ letters or the content of the options.
- For open-ended questions:
  * Score 1 if the prediction matches the answer semantically, it can be in different format.
  * Score 0 for partially correct answers or answers with extra incorrect information, even if the reasoning process is correct.
- Ignore minor differences in formatting, capitalization, or spacing since the model may explain in a different way.
- Treat numerical answers as correct if they match within reasonable precision
- For questions requiring units, both value and unit must be correct

# Strict Output format
1 or 0
```

Table 7. The prompt that used in evaluation.

ColdStart Data Samples



Question	As shown in the figure, line EF passes through the intersection point O of the diagonals of the parallelogram and intersects AD and BC at points E and F, respectively. If the area of the parallelogram is 12, the sum of the areas of triangles AOE and BOF equals ( )A. 2 B. 3 C. 4 D. Cannot be determined Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.
Answer	<think>In a parallelogram, the diagonals bisect each other, so point O is the midpoint of both diagonals AC and BD. Also, since EF passes through O and intersects AD and BC at E and F respectively, we can use symmetry and properties of parallelograms. Because O is the center of the parallelogram, any line through O divides the parallelogram into two regions of equal area. Moreover, triangles AOE and COF are congruent (by ASA or SAS, since AO = OC, ∠AOE = ∠COF, and ∠OAE = ∠OCF due to parallel lines and transversal), and similarly, triangles BOF and DOE are congruent.Therefore, the area of triangle AOE equals the area of triangle COF, and the area of triangle BOF equals the area of triangle DOE.The total area of the parallelogram is 12. The diagonals divide it into 4 triangles of equal area: AOB, BOC, COD, DOA — each of area 3.Now, consider the two shaded triangles: AOE and BOF. Since EF passes through O, and by symmetry, the sum of the areas of AOE and BOF is equal to the area of triangle AOB (which is one of the four equal-area triangles formed by the diagonals).Why? Because:- Area(AOE) + Area(BOF) = Area(AOB)This is because the parts not included in AOE and BOF (namely, EOD and FOC) are congruent to AOE and BOF respectively, so swapping them doesn't change the total.Alternatively, since the entire parallelogram is divided into 4 triangles of area 3 each, and AOE + BOF together make up exactly one of those triangles (AOB), their combined area is 3.Thus, the answer is 3.</think><answer>B. 3</answer>

Table 8. An example of the coldstart data sample.

ColdStart Data Samples

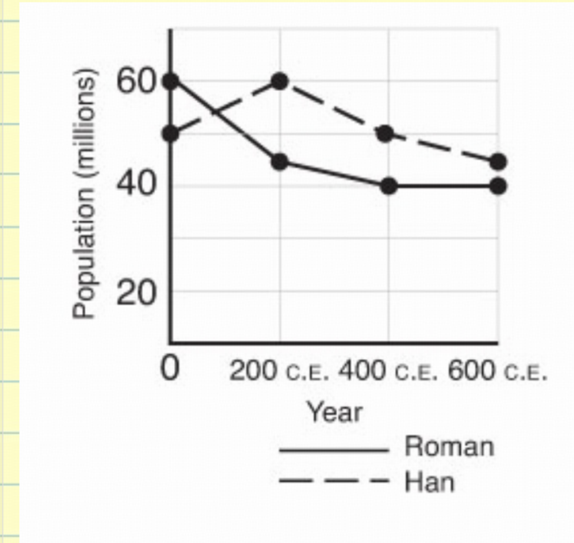
Question: What is the man doing with the beans?A. cookingB. eating themC. countingD. selling themAnswer with the option's letter from the given choices directly. Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.

Answer: <think>The image shows a man sitting beside a cart loaded with green beans in what appears to be a busy street market. He is handling the beans and seems to be interacting with potential customers, which is typical behavior for a vendor. There is no visible cooking equipment, nor is he eating or counting the beans. The context strongly suggests he is selling them.</think><answer>D</answer>

Table 9. An example of the coldstart data sample.

Qualitative Data Sample

No. 25: Which of the following best explains the overall trend shown in the [image 1] ?



Options:

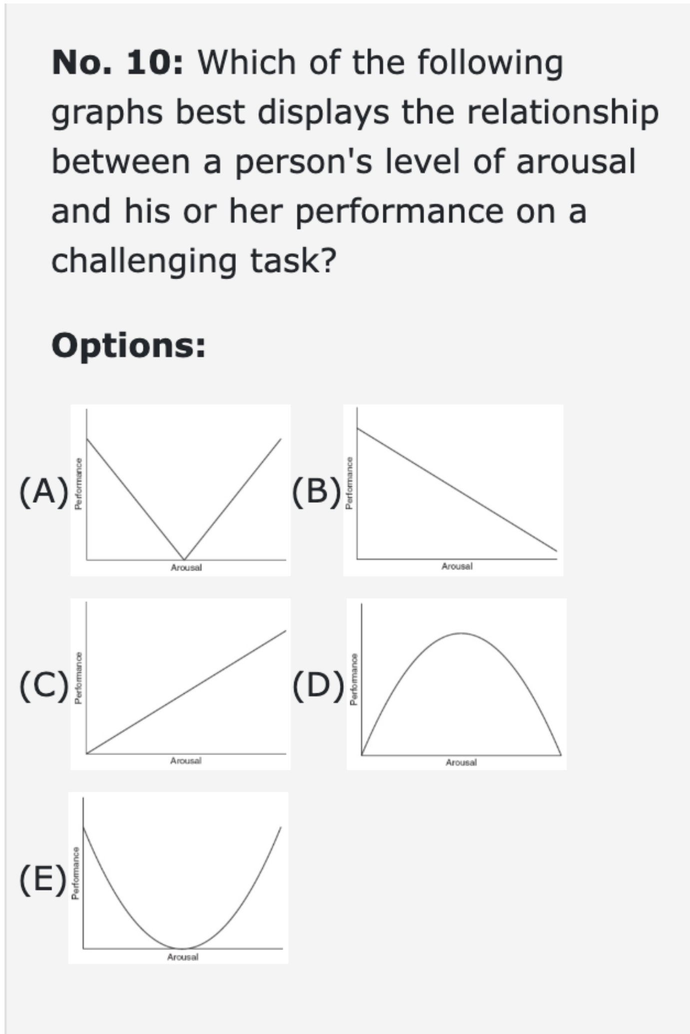
(A)Political instability leading to population decline
(B)The spread of pathogens across the Silk Road
(C)Development of new trade routes
(D)Climate change affecting the Silk Road
(E)Migrations to areas of Central Asia for resettlement
(F)Technological advancements in transportation
(G)Invasions by Mongol tribes
(H)Large-scale famine due to crop failures
(I)Economic prosperity and population growth
(J)Rise of religious conflicts along the Silk Road

Previous Next 1 / 170 Go

Question	A. Political instability leading to population decline B. The spread of pathogens across the Silk Road C. Development of new trade routes D. Climate change affecting the Silk Road E. Migrations to areas of Central Asia for resettlement F. Technological advancements in transportationG. Invasions by Mongol tribes H. Large-scale famine due to crop failures I. Economic prosperity and population growth J. Rise of religious conflicts along the Silk Road Answer the question in the image.
Answer	B
Ours	<think>The graph shows population trends for Roman and Han populations over time, with a general decline in population for both over centuries. This decline is most plausibly explained by widespread disease, as the Silk Road facilitated the spread of pathogens across regions, leading to population reductions. While political instability or other factors could contribute, the most direct and historically supported cause for such a widespread population decline across empires is the spread of pathogens.</think><answer>B. The spread of pathogens across the Silk Road</answer>
Qwen2.5-VL-7B-Instruct	<think>The graph shows a general decline in population for both the Roman and Han empires over the period from 200 C.E. to 600 C.E. This trend is not consistent with economic prosperity, technological advancements, or large-scale famine, which would typically lead to population growth. The spread of pathogens across the Silk Road could potentially cause a decline in population, but the graph does not provide specific information about the cause of the decline. The rise of religious conflicts along the Silk Road could also lead to population decline, but the graph does not provide specific information about the cause of the decline. The development of new trade routes and the rise of religious conflicts along the Silk Road are not directly related to the population decline shown in the graph. The graph does not provide specific information about the cause of the decline, but it is reasonable to infer that the decline in population could be due to political instability, which could lead to a decline in the overall population.</think><answer>A. Political instability leading to population decline</answer>

Table 10. An example of our model inference result

Qualitative Data Sample



Question	A. <image 3> B. <image 2> C. <image 4> D. <image 5> E. <image 1> Answer the question in the image.
Answer	D
Ours	<think> The relationship between arousal and performance on a challenging task is typically described by the Yerkes-Dodson Law, which states that performance increases with arousal up to a point, after which further increases in arousal lead to a decline in performance. This forms a bell-shaped curve 2014 peak performance at moderate arousal, with performance decreasing as arousal rises too high or falls too low. Looking at the options: - (A) Inverted V: performance improves with low arousal, then drops 2014 doesn2019t match. - (B) Straight line decreasing: performance always decreases with arousal 2014 too linear and wrong direction. - (C) Straight line increasing: performance always increases with arousal 2014 incorrect. - (D) Bell curve: performance peaks at moderate arousal 2014 matches Yerkes-Dodson Law. - (E) U-shape: performance improves with low arousal, then drops 2014 opposite of what2019s expected. Thus, (D) best represents the relationship. The question asks to choose from A-E, and (D) is labeled as image 5 in the options mapping. So, the correct answer is D. </think><answer>D. <image 5></answer>
Qwen2.5-VL-7B-Instruct	<think>The relationship between a person's level of arousal and their performance on a challenging task is often described by the Yerkes-Dodson Law, which suggests that performance increases with arousal up to a certain point and then decreases. This relationship is typically represented by a bell-shaped curve, indicating an optimal level of arousal for peak performance.</think><answer>E</answer>

Table 11. An example of our model inference result