MM-OPD: TOWARDS ONE MORE BOTTLENECK BETWEEN PERCEPTION AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent multimodal large language models (MLLMs) advance visual reasoning by strengthening both perception and reasoning, implicitly assuming a process that transitions seamlessly from perception to reasoning. However, we observe a counterintuitive phenomenon that challenges this assumption: holding the model, question, and decoding fixed, we replace images with their caption or code representations (*symbolic views*), which seems to be redundant given the clear image structures, but the performance surprisingly improves by 10.2% to 23.6% across model scales and datasets. We term this performance gap as the ***Symbolic Visual Gap*** and then take a closer look at it. Through experiments, we find that although the visual evidence can already appear in the reasoning trace for the image-input model, the symbolic-view-input model shows much higher attention to the correct evidence than the image-input model. This suggests that despite good capabilities from current works in perception and reasoning themselves, another bottleneck exists **between** perception and reasoning in selecting perceived visual information as appropriate evidence for subsequent reasoning. To handle this bottleneck, since the symbolic view steers attention toward correct evidence and is readily obtained at scale, it provides supervision for evidence selection without manually labeled evidence. Building on this, we introduce **MM-OPD**, a multimodal on-policy self-distillation framework for *symbolic-to-visual correction* that transfers guidance from symbolic-conditioned behavior to the image-conditioned policy through residual token-level targets, steering the model toward correct visual evidence. Experiments across benchmarks and model scales show that **MM-OPD** improves a broad range of multimodal abilities, with gains in visual perception, chart and document understanding, mathematical reasoning, and general VQA.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have shown promising capabilities in visual reasoning. Some efforts emphasize access to fine-grained visual information (Liu et al., 2025; Su et al., 2026), while others strengthen multi-step mathematical and logical reasoning through post-training (Shao et al., 2024; Hong et al., 2025), assuming an ideal visual reasoning process that seamlessly transitions from perception to reasoning.

However, in this paper, we observe a counterintuitive phenomenon contradicting the seamless perception-reasoning transition. As in Fig. 1, we use the generating code (or descriptive caption, which we term the *symbolic view*) to replace the question image as the model input, with the model, question, and the decoding fixed, and measure the reasoning performance. Since the input image is simple enough for perception, the image is ideally itself easier to understand than the generating code, making the replacement redundant. Surprisingly, such a seemingly redundant operation significantly improves performance by 10.2% to 23.6% across model scales and datasets (Fig. 1, right). This phenomenon makes us question the widely assumed seamless transition from perception to reasoning in visual reasoning: *do we need another symbolic-view step during visual reasoning?*

To answer this question, we refer to this discrepancy between image-conditioned reasoning and symbol-conditioned reasoning as the ***Symbolic Visual Gap*** and take a step closer to it. We first analyze the visual information expressed in the generated reasoning traces for image-input models. We find that the correct visual evidence can already appear in the trace, indicating that the model can al-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

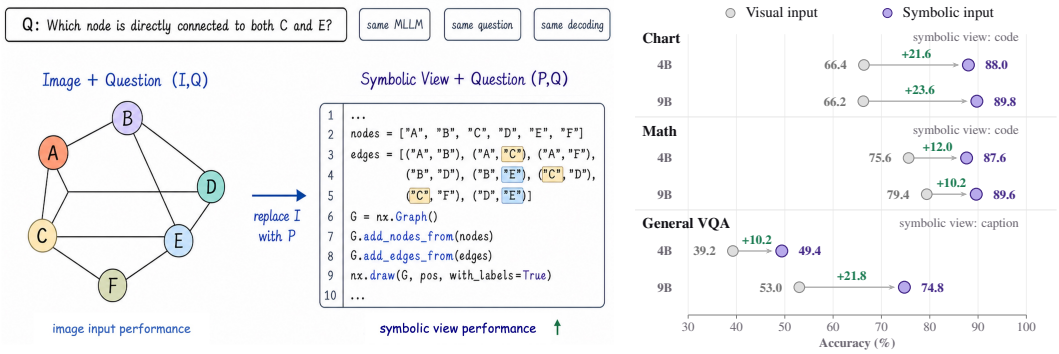


Figure 1: **(Left)** A case of replacing the question image with its symbolic view, under the same model, question, and decoding procedure. **(Right)** The *Symbolic Visual Gap*: symbolic views outperform visual inputs on every model and dataset pair from 4B to 9B.

ready correctly perceive the visual information. However, we still find that the model tends to adopt distracting evidence (visual evidence leading to incorrect reasoning and answers) for its subsequent reasoning, showing a disadvantage in visual evidence selection. To verify it, we then quantitatively measure the answer-to-evidence attention between the image input and the symbolic-view input. We find that answer-to-evidence attention favors distracting evidence given the image input, and correct evidence given the symbolic-view input, verifying the difficulty in selecting correct evidence for the image-input model. In other words, despite good capabilities from current works in perception and reasoning themselves, **another bottleneck exists between perception and reasoning in selecting perceived visual information as appropriate evidence for subsequent reasoning**.

To handle this bottleneck, based on the above analysis, as the symbolic view elicits the model to assign answer-to-evidence attention to correct visual evidence, this view naturally offers a direction for guiding the image-input model for evidence selection without manually labeling correct evidence. Moreover, since the symbolic view (generating code or descriptive caption) is easily obtained automatically, it is scalable for different data sizes. Therefore, we propose to take the symbolic view as auxiliary information and introduce **MM-OPD**, a multimodal on-policy self-distillation framework for correcting visual-evidence selection. **MM-OPD** takes the symbolic-view-input model as the teacher and uses the image-input model as the student, and then provides token-level supervision along trajectories generated by the student to guide the student to learn the evidence selection from the self-teacher. During inference, since the symbolic view is not available, only the image-input model, with the symbolic visual gap largely reduced, will be used for visual reasoning. **MM-OPD** improves a broad range of multimodal abilities rather than a single targeted capability. Experiments across diverse benchmarks show consistent gains in visual perception, chart and document understanding, mathematical reasoning, and general multimodal reasoning. Our contributions are as follows:

- We identify the *Symbolic Visual Gap*: with the model, question, and decoding fixed, replacing a visual input with its symbolic view improves accuracy by 10.2 to 23.6 percentage points across model scales and visual domains.
- We identify visual evidence selection as another critical bottleneck between perception and reasoning: correct visual information can appear in the generated trace without guiding the derivation, while image and symbolic conditioning exhibit opposite answer-to-evidence attention patterns.
- Based on this insight, we introduce **MM-OPD**, a multimodal on-policy self-distillation framework, to transfer guidance from symbolic-conditioned behavior to the image-conditioned policy through residual token-level targets to help the model draw on correct visual evidence.
- Comprehensive experiments validate the effectiveness of **MM-OPD**, demonstrating consistent improvements across model scales and diverse multimodal reasoning tasks.

2 Delve into the Symbolic Visual Gap

The counterintuitive Symbolic Visual Gap inspires us to examine the assumed seamless transition from perception to reasoning more closely. In an ideal visual reasoning process, perceived

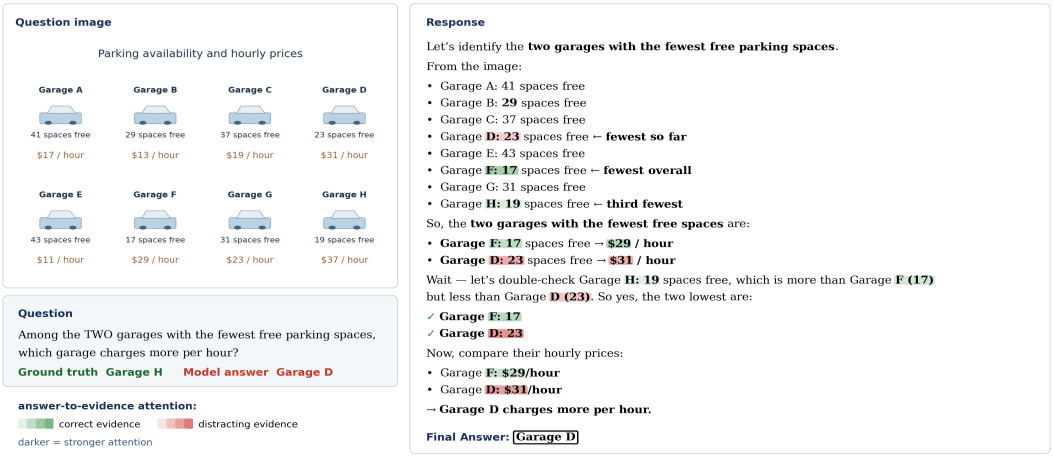


Figure 2: A case study. Task-relevant visual information is correctly perceived and verbalized, yet distracting evidence receives more attention and is selected as the premise for subsequent reasoning.

information should become appropriate evidence for subsequent reasoning. We therefore examine (i) whether task-relevant visual information is correctly perceived and represented as language-accessible evidence, and (ii) whether subsequent reasoning draws on the correct evidence.

2.1 From Visual Perception to Evidence Selection

When input an image into the MLLM (i.e., the image-input model), in the model-generated reasoning, some tokens verbalize information obtained from the input image that may be used to answer the question. We refer to these as *visual evidence tokens*. For a given question, *correct evidence tokens* express visual information that supports a correct answer, whereas *distracting evidence tokens* express competing information that can mislead subsequent reasoning or answer generation. To distinguish visual perception from evidence use, we examine whether correct visual information already appears in the generated trace. Its explicit presence indicates that this information is available in a language-accessible form. We then inspect which evidence is used in subsequent reasoning and compare answer-to-evidence attention for correct and distracting evidence, thereby characterizing how strongly answer generation is associated with each type of candidate evidence.

Figure 2 illustrates that correctly perceived visual information may not be selected as appropriate evidence for subsequent reasoning. The relevant values and their ordering are explicitly stated in the trace, yet the model still selects distracting evidence as a premise. The subsequent comparison is valid given that premise. Distracting evidence also receives more attention, including the evidence used in the derivation. Thus, having relevant visual information available in the trace does not ensure that it is used appropriately to answer the question. Additional cases are provided in Appendix E.

Finding 1 Correctly perceiving and verbalizing task-relevant visual information does not ensure correct evidence use. The model may select distracting evidence for subsequent reasoning.

2.2 Answer-to-Evidence Attention under Image and Symbolic Conditions

We next quantitatively examine whether the attention preference observed in the case study extends beyond this example and how it differs between image and symbolic conditioning. Specifically, we sample 200 examples from each of three task families (general VQA, chart reasoning, and mathematical reasoning), yielding 600 examples in total. Using Qwen3.5-4B, we measure at each full-attention layer the head-averaged attention from the answer token to correct and distracting evidence tokens under image and symbolic conditioning, and report both layer-wise and overall results.

i) The two conditions exhibit opposite evidence orderings across depth. As in Fig. 3(a,b), distracting evidence receives more attention than correct evidence at every observed layer under image conditioning. Symbolic conditioning consistently exhibits the opposite ordering. This reversal persists across all observed layers and is therefore not driven by an isolated layer.

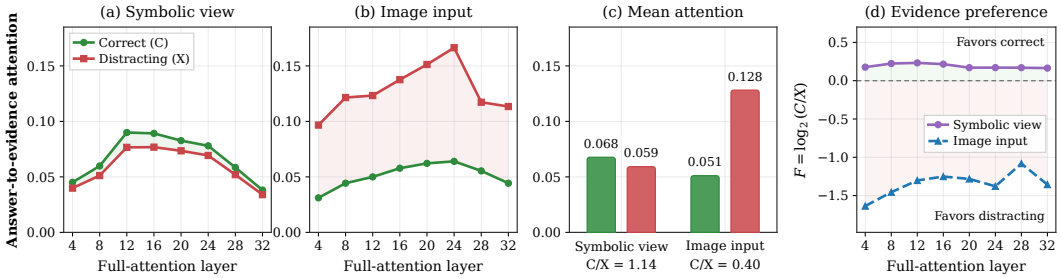


Figure 3: **Answer-to-evidence attention under image and symbolic conditioning.** (a,b) Layer-wise results at the observed full-attention layers. (c) Overall mean attention to correct and distracting evidence, together with their ratio. (d) The signed score $F = \log_2(C/X)$; positive and negative values indicate stronger association with correct and distracting evidence, respectively.

ii) The overall means and evidence ratios preserve the same trend. The overall statistics in Fig. 3(c) agree with the layer-wise results. Let C and X denote the mean attention to correct and distracting evidence, respectively. We define their correct-to-distracting ratio as $R = \frac{C}{X}$, where $R > 1.0$ and $R < 1.0$ indicate stronger association with correct and distracting evidence, respectively. The ratio is above 1.0 under symbolic conditioning and below 1.0 under image conditioning.

iii) A signed score summarizes the reversal. For a signed summary of the same relative attention, we define $F = \log_2 R = \log_2(C/X)$. Positive values indicate a stronger association with correct evidence, whereas negative values indicate a stronger association with distracting evidence. As shown in Fig. 3(d), F is positive at every observed layer under symbolic conditioning and negative at every observed layer under image conditioning. The overall scores retain the same signs.

Together with the case study, the consistent attention preference for distracting evidence under image conditioning points to a difficulty in using perceived information as evidence for reasoning.

Finding 2 Visual evidence selection is one more bottleneck between perception and reasoning. Beyond perception and reasoning, effective multimodal capability also depends on whether perceived information is selected as appropriate evidence for subsequent reasoning.

2.3 Summary and Discussion

The above analyses challenge the assumed seamless transition from perception to reasoning. Relevant visual information can already appear in the generated trace, yet the model may still select distracting evidence for subsequent reasoning. In other words, perceiving the required information does not ensure that it becomes appropriate evidence for answering the question.

To address this bottleneck, the comparison with symbolic views offers a natural direction. With the model and question fixed, symbolic conditioning elicits stronger performance and an attention preference for correct evidence. This contrast suggests that the behavior elicited by symbolic views can guide evidence selection under image conditioning. Moreover, symbolic views, instantiated as generating code or descriptive captions, can be obtained automatically at scale, providing a scalable source of guidance without manually labeling correct evidence.

Insight Symbolic Visual Gap reveals a missing link between perception and reasoning: visual information may be perceived without being correctly selected. Symbolic views elicit better evidence selection that can guide image-conditioned reasoning without manual evidence labeling.

3 Method

The above analysis reveals visual evidence selection as one more bottleneck between perception and reasoning. It also shows that symbolic views steer answer-to-evidence attention toward correct visual evidence. Since these views can be obtained automatically at scale, this finding motivates us to introduce **MM-OPD**, a multimodal on-policy self-distillation framework for symbolic-to-visual

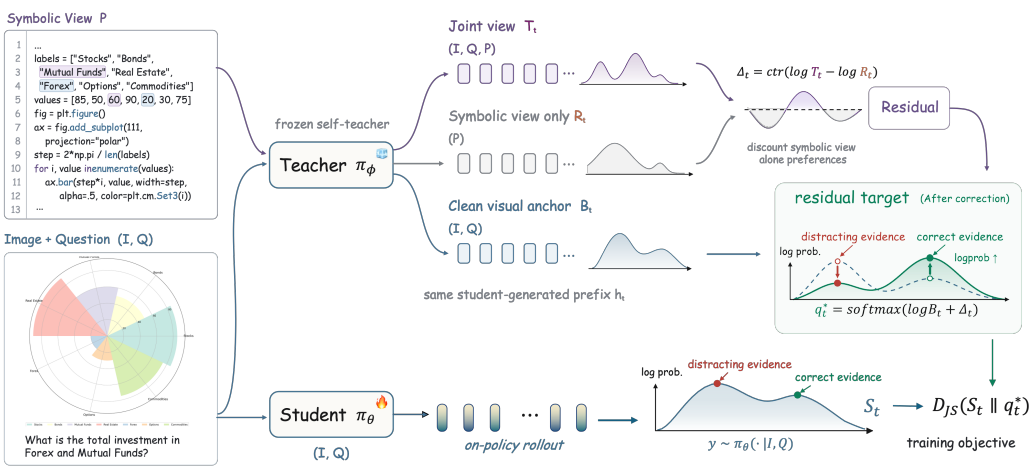


Figure 4: Overview of **MM-OPD**, a multimodal OPSD framework for symbolic-to-visual correction. Symbolic views provide privileged supervision, while residual targets discount token preferences induced by these views alone and guide image-conditioned reasoning toward correct visual evidence.

correction. A frozen self-teacher uses additional symbolic information to supervise the student along student-generated trajectories. Residual token-level targets reduce the influence of token preferences induced by the symbolic view alone and refine image-conditioned predictions. Figure 4 illustrates the framework, with training pseudocode in Appendix A.

3.1 Symbolic-to-Visual Correction

Given the curated dataset $\mathcal{D} = \{(I_i, Q_i, P_i)\}_{i=1}^N$, where P_i is privileged symbolic information, we initialize a trainable student π_θ and a frozen self-teacher π_ϕ from the same pretrained MLLM checkpoint. For each $(I, Q, P) \sim \mathcal{D}$, the image-conditioned student generates an on-policy response:

$$y = (y_1, \ldots, y_{|y|}) \sim \pi_\theta(\cdot \mid I, Q), \qquad h_t = y_{<t}, \tag{1}$$

where h_t is the student-generated prefix at position t. Let $\mathcal{V}$ denote the model vocabulary and $v \in \mathcal{V}$ a candidate next token. We evaluate four distributions on the same prefix:

$$\begin{aligned} S_t(v) &= \pi_\theta(v \mid I, Q, h_t), \quad T_t(v) = \pi_\phi(v \mid I, Q, P, h_t), \\ B_t(v) &= \pi_\phi(v \mid I, Q, h_t), \quad R_t(v) = \pi_\phi(v \mid P, h_t). \end{aligned} \tag{2}$$

Here, S_t is the current student prediction. The other three distributions are obtained from the single frozen self-teacher under different conditioning views: T_t combines the image, question, and privileged symbolic information; B_t is the clean image-conditioned prediction; and R_t estimates the preference induced by P without direct access to the image or question. Their differences reflect the conditioning view rather than different generated trajectories.

Residual target construction Directly distilling T_t would conflate the task-conditioned benefit of P with token preferences that P can induce on its own. **MM-OPD** therefore uses R_t as a privilege-only reference and contrasts it with T_t. For the same token, the log-ratio $\log T_t(v) - \log R_t(v)$ discounts the preference explained by P alone, leaving a correction associated with interpreting P in the full image–question context.

For tractable token-level distillation, the privileged teacher defines a shared candidate set $\mathcal{K}_t = \mathrm{Top}_K(T_t)$ containing its K most probable tokens; probabilities outside this set are later aggregated into a tail bucket. We remove the vocabulary-wide offset from the log-ratio:

$$\widetilde{d}_t(v) = \log T_t(v) - \log R_t(v) - \mathbb{E}_{u \in \mathcal{V}} \left[\log T_t(u) - \log R_t(u)\right]. \tag{3}$$

The expectation is uniform over $\mathcal{V}$. Subtracting it removes a token-independent shift, which carries no information about the relative preference among candidate tokens. Thus, $\widetilde{d}_t(v)$ measures how the full task context changes the relative preference for v beyond what is induced by P alone.

Rather than replacing the image-conditioned prediction with the privileged teacher, we use this residual to correct the clean anchor B_t:

$$\widehat{q}_t = \mathrm{softmax}_{\mathcal{K}_t}\left(\log B_t + \gamma \tanh\left(\frac{\widetilde{d}_t}{\gamma}\right)\right). \tag{4}$$

Adding the residual in log-probability space reweights the token preferences of B_t, so privileged information acts as a correction to the prediction made from the actual inference input. The saturation function remains approximately linear for moderate residuals while suppressing extreme log-ratios, with $\gamma > 0$ controlling its scale.

Finally, let $M_{B,t} = \sum_{v\in\mathcal{K}_t} B_t(v)$ be the probability mass that the clean anchor assigns to the selected support. We preserve this mass and distribute it according to $\widehat{q}_t$:

$$q_t^\star(v) = M_{B,t}\widehat{q}_t(v), \qquad v \in \mathcal{K}_t. \tag{5}$$

Writing tail for all tokens outside $\mathcal{K}_t$, we retain the remaining anchor mass:

$$q_t^\star(\mathrm{tail}) = 1 - M_{B,t} = B_t(\mathrm{tail}). \tag{6}$$

Hence, the residual changes only the relative probabilities within the teacher-selected support; the total Top-K and tail masses remain grounded in the clean image-conditioned prediction.

3.2 Training Objective

To match the support of $q_t^\star$, we compress the student prediction S_t onto $\mathcal{K}_t \cup \{\mathrm{tail}\}$ by retaining $\widetilde{S}_t(v) = S_t(v)$ for $v \in \mathcal{K}_t$ and aggregating the remaining mass as $\widetilde{S}_t(\mathrm{tail}) = 1 - \sum_{u\in\mathcal{K}_t} S_t(u)$. Both $\widetilde{S}_t$ and $q_t^\star$ are therefore normalized on the same support. We minimize their symmetric Jensen–Shannon divergence over the student's on-policy trajectories:

$$\mathcal{L}_{\text{MM-OPD}}(\theta) = \mathbb{E}_{\substack{(I,Q,P)\sim\mathcal{D}\\ y\sim\pi_\theta(\cdot|I,Q)}}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{JS}}\left(\widetilde{S}_t \,\middle\|\, q_t^\star\right)\right]. \tag{7}$$

The teacher-side distributions T_t, B_t, and R_t are detached, and $q_t^\star$ is treated as a fixed target, so gradients flow only through the image-conditioned student. Because each trajectory is generated by the student, the objective provides token-level supervision on the state distribution encountered at inference time.

3.3 Training Dataset Curation

We construct $\mathcal{D} = \{(I_i, Q_i, P_i)\}_{i=1}^N$, where P_i is privileged symbolic information, instantiated as a caption for a natural image or executable code for a structured visual. We sample paired examples from V-Interaction (Qiao et al., 2025b), Mulberry (Yao et al., 2026), and ChartX (Xia et al., 2025), and curate them in three stages. *Cross-view validation* discards caption–image pairs that are semantically inconsistent and code that fails to execute. *Symbolic-view normalization* reorganizes the dataset-provided captions and code into a consistent format. *Symbolic-view utility filtering* keeps only examples whose symbolic view does not lower $\mathrm{Pass@8}$. The resulting dataset contains 32K examples spanning general VQA, mathematical diagrams, and charts; see Appendix B for details.

4 Experiments

4.1 Experimental Setup

Models and data. We apply **`MM-OPD`** to Qwen3.5-4B and Qwen3.5-9B (Qwen Team, 2026), both trained on our curated MMOPD-32K, which pairs each image–question example with a symbolic view: a caption for a natural image or executable code for a chart or mathematical diagram. The symbolic view is used only during training and in the controlled gap analysis. Full training and inference configurations are given in Appendix D.

Benchmarks and evaluation. We evaluate four capability groups: *general VQA* (Chen et al., 2024; XAI, 2024; Liu et al., 2024a; Cheng et al., 2025), *multimodal reasoning* (Lu et al., 2024; Zhang et al.,

Table 1: **Main results across diverse multimodal benchmarks.** The best result in is shown in **bold** and the second best is underlined.

Category	Benchmark	Qwen3.5-4B							Qwen3.5-9B						
		Baseline	SFT	GRPO	GSPO	DAPO	OPSD	MM-OPD	Baseline	SFT	GRPO	GSPO	DAPO	OPSD	MM-OPD
General VQA	MMStar	72.9	70.1	73.8	72.1	68.7	72.8	**74.6**	74.9	70.5	73.4	74.5	73.0	75.4	**75.9**
	RealWorldQA	75.7	75.4	**77.8**	76.5	74.2	76.1	77.4	78.0	**78.3**	76.1	76.2	76.1	77.9	77.5
	MMBench v1.1	85.7	83.5	86.5	86.3	84.8	85.2	**86.6**	87.2	84.4	86.1	85.7	86.0	87.2	**87.7**
	SimpleVQA	50.7	49.9	51.2	47.5	**52.2**	48.4	51.7	51.6	49.2	52.4	**53.3**	52.9	53.2	53.2
Reasoning Tasks	MathVista$_{mini}$	81.4	75.9	81.8	80.5	75.7	81.8	**82.2**	82.9	77.9	83.2	82.9	83.4	83.0	**84.1**
	MathVerse$_{mini}$	53.9	50.2	53.3	53.9	54.3	53.4	**54.5**	54.1	52.0	47.9	49.9	54.9	52.9	**57.9**
	DynaMath	72.4	70.0	71.9	71.5	70.3	70.9	**74.0**	75.0	72.4	75.8	73.4	74.4	76.1	**77.7**
	We-Math	71.1	69.0	70.4	70.6	70.4	71.1	**74.4**	73.0	66.7	74.5	73.5	70.2	74.2	**76.5**
	LogicVista	66.7	61.6	63.1	66.0	65.6	68.0	**69.4**	67.3	64.7	65.5	67.1	63.1	70.2	**73.4**
	MMMU	72.6	68.7	71.3	71.7	70.5	72.2	**73.8**	74.7	70.0	72.6	74.7	71.1	74.1	**75.8**
	MMMU-Pro	57.6	53.6	56.3	57.3	55.9	**59.2**	58.5	62.7	55.1	59.5	59.9	58.0	64.1	**64.7**
Perception Tasks	V* Bench	81.7	76.4	83.8	84.3	80.1	84.3	**88.5**	82.2	84.3	83.8	84.3	82.2	**86.9**	84.8
	HR-Bench 4K	81.5	80.4	83.6	81.6	78.0	80.3	**85.0**	82.7	80.9	82.3	82.4	80.0	81.3	**84.0**
	HR-Bench 8K	75.1	72.4	76.1	76.9	73.3	76.0	**80.5**	77.6	75.8	78.0	76.9	77.5	**79.6**	78.4
Document Understanding	ChartQA	84.2	80.4	83.9	76.4	80.2	84.4	**85.1**	85.4	81.7	84.6	84.8	83.6	84.0	**93.7**
	CharXiv (DQ)	88.0	81.3	88.6	85.6	87.6	**88.9**	88.0	91.1	85.9	89.8	88.0	90.0	90.7	**91.8**
	CharXiv (RQ)	65.9	62.8	67.6	62.5	62.2	64.5	**68.8**	69.0	63.1	68.6	67.7	68.2	68.6	**71.2**
	DocVQA	95.1	94.2	95.3	94.7	94.9	95.0	**95.5**	94.9	**95.4**	94.9	94.9	95.0	95.0	**95.4**
	OCRBench	87.1	81.3	87.4	86.5	86.4	87.6	**87.8**	85.9	84.0	**87.3**	86.4	86.7	86.0	87.0
	OCRBench v2 (EN)	53.9	48.9	53.8	55.0	52.0	52.5	**55.1**	56.3	47.5	56.5	56.4	56.8	56.3	**57.2**
Total	**Average**	73.7	70.3	73.9	72.9	71.9	73.6	**75.6**	75.3	72.0	74.6	74.6	74.2	75.8	**77.4**

2024; Zou et al., 2025; Qiao et al., 2025a; Xiao et al., 2024; Yue et al., 2024; 2025), *fine-grained perception* (Wu & Xie, 2024; Wang et al., 2025), and *document and chart understanding* (Masry et al., 2022; Wang et al., 2024; Mathew et al., 2021; Liu et al., 2024b; Fu et al., 2026). For a fair comparison, all checkpoints are evaluated with VLMEvalKit (Duan et al., 2024) under a unified protocol, and Qwen3-32B is used as the judge model for all LLM-judged benchmarks.

Compared methods. All baselines start from the same Qwen3.5 checkpoint and share the MMOPD-32K training set. **Base** is the unchanged model; **SFT** performs teacher-forced fine-tuning on reference responses (Ouyang et al., 2022); **GRPO** (Shao et al., 2024), **GSPO** (Zheng et al., 2025), and **DAPO** (Yu et al., 2026a) are RLVR baselines using the same Qwen3-32B judge; and **Vanilla OPSD** (Zhao et al., 2026) applies on-policy self-distillation with the reference answer as the teacher's privileged information, sharing our on-policy sampling, frozen self-teacher, and top-K support but differing in both the privileged signal and the target construction.

4.2 Main Results

We first compare **MM-OPD** with controlled post-training baselines across all benchmarks.

Broad and consistent gains across capabilities. As shown in Table 1, **MM-OPD** achieves the highest average score at both model scales, improving over the Qwen3.5-4B and Qwen3.5-9B baselines by 1.9 and 2.1 points, respectively. The gains span all four capability groups, including general VQA, multimodal reasoning, fine-grained perception, and document understanding. These results show that **MM-OPD** improves multimodal capabilities broadly rather than targeting a single task family.

Stronger overall performance than SFT and RLVR. Under the same training data and backbones, SFT lowers the average score at both scales, while GRPO, GSPO, and DAPO produce uneven changes across benchmarks. In contrast, **MM-OPD** surpasses the best average among these RLVR baselines by 1.7 points at 4B and 2.8 points at 9B. Its gains are therefore both stronger overall and more broadly distributed across capabilities.

Benefits over vanilla OPSD. Vanilla OPSD shares the on-policy sampling procedure and frozen self-teacher with **MM-OPD**, but conditions the teacher on the reference answer and directly distills its prediction. In contrast, **MM-OPD** uses the symbolic view and constructs a residual correction. It improves the average over vanilla OPSD by 2.0 points at 4B and 1.6 points at 9B, including a 5.0-

point gain on MathVerse-mini at 9B. This comparison supports the combined benefit of symbolic supervision and residual correction in multimodal on-policy self-distillation.

4.3 Ablation Study

We conduct all ablations with Qwen3.5-4B on six benchmarks covering general VQA, multimodal reasoning, and fine-grained perception. Each experiment varies one factor while keeping the remaining settings at their defaults.

Table 2: **Effect of the distillation objective on Qwen3.5-4B.** Forward and reverse KL denote $D_{\mathrm{KL}}(q_t^\star \| \widetilde{S}_t)$ and $D_{\mathrm{KL}}(\widetilde{S}_t \| q_t^\star)$, respectively. The residual target construction is unchanged.

Divergence	V* Bench	HR-Bench 4K	HR-Bench 8K	MMStar	MathVista$_{mini}$	LogicVista	Average
Forward KL	84.3	84.8	79.5	74.4	82.0	67.6	78.8
Reverse KL	86.4	**85.3**	79.6	74.5	**82.7**	**70.0**	79.8
JSD	**88.5**	85.0	**80.5**	**74.6**	82.2	69.4	**80.0**

Distillation objective. As shown in Table 2, we compare forward KL, reverse KL, and JSD as the token-wise divergence objective. All three objectives improve over the baseline on every benchmark, showing that the benefits of **MM-OPD** are not specific to a single divergence function. JSD and reverse KL each lead on three benchmarks, indicating different preferences across tasks. JSD achieves the highest overall average and is therefore adopted as our default.

Table 3: **Effect of the teacher update rule on Qwen3.5-4B.**

Teacher Update	V* Bench	HR-Bench 4K	HR-Bench 8K	MMStar	MathVista$_{mini}$	LogicVista	Average
EMA	81.1	**85.8**	79.8	**75.0**	81.6	67.1	78.4
Frozen	**88.5**	85.0	**80.5**	74.6	**82.2**	**69.4**	**80.0**

Teacher update rule. We examine how updating the self-teacher affects distillation (Table 3). Both EMA and frozen teachers improve over the baseline on average. EMA performs better on HR-Bench 4K and MMStar, whereas the frozen teacher leads on the remaining four benchmarks and achieves a 1.6-point higher average. We therefore keep the teacher parameters frozen at initialization.

Table 4: **Effect of the maximum on-policy response length.**

Response Length	V* Bench	HR-Bench 4K	HR-Bench 8K	MMStar	MathVista$_{mini}$	LogicVista	Average
1024	84.3	83.5	78.6	75.0	81.4	66.2	78.2
2048	84.8	84.1	79.1	**75.4**	81.7	66.4	78.6
4096	86.4	84.2	79.9	73.9	82.1	**69.6**	79.4
8192	**88.5**	85.0	**80.5**	74.6	82.2	69.4	**80.0**
16384	86.9	**85.2**	**80.5**	74.3	**82.7**	68.1	79.6

Maximum response length. As shown in Table 4, we vary the maximum length of on-policy training responses from 1024 to 16384 tokens. Different benchmarks favor different length limits, and increasing the limit does not consistently improve performance on every task. Nevertheless, all tested limits outperform the baseline on average, demonstrating robust gains across the evaluated range. Average performance peaks at 8192; we therefore adopt 8192 as the default.

Table 5: **Effect of the number of on-policy rollouts.**

Rollout Number	V* Bench	HR-Bench 4K	HR-Bench 8K	MMStar	MathVista$_{mini}$	LogicVista	Average
1	84.3	83.5	**81.0**	74.1	**82.9**	68.9	79.1
2	**88.5**	85.0	80.5	**74.6**	82.2	**69.4**	**80.0**
4	86.4	**85.5**	79.3	74.4	82.2	69.1	79.5
8	86.9	85.0	79.1	74.5	82.0	68.7	79.4

Number of rollouts. Keeping all other settings at their default values, we vary the number of on-policy trajectories per prompt (Table 5). All tested rollout counts outperform the baseline on every benchmark. Increasing the count from one to two improves the average by 0.9 points, while further increases to four or eight yield slightly lower averages. The gains therefore persist across rollout counts, with two rollouts per prompt achieving the highest average score and serving as our default.

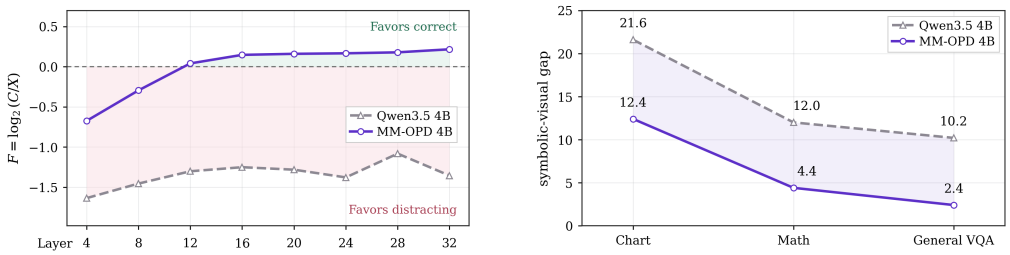


Figure 5: **Effectiveness of Symbolic-to-Visual Correction.** (Left) Layer-wise answer-to-evidence preference $F = \log_2(C/X)$ under image conditioning, where positive values favor correct evidence. (Right) The Symbolic Visual Gap before and after training across three task families.

4.4 Effectiveness of Symbolic-to-Visual Correction

We examine whether **MM-OPD** transfers the advantage of symbolic views to image-conditioned reasoning using two complementary diagnostics on Qwen3.5-4B: answer-to-evidence preference and the Symbolic Visual Gap. Our experiment show that: **1) Evidence preference shifts toward correct evidence.** As shown in Fig. 5 (left), the base model favors distracting evidence at every observed full-attention layer. After training, **MM-OPD** increases the relative attention to correct evidence across all layers, with correct evidence becoming dominant from layer 12 onward. This shift indicates a stronger association between answer generation and evidence supporting the correct derivation. **2)The Symbolic Visual Gap narrows across tasks.** Using the base model's symbolic-conditioned accuracy as a fixed reference, **MM-OPD** reduces the gap on Chart, Math, and General VQA, lowering the unweighted mean gap by 56.2% (Fig. 5, right). Together, the two diagnostics show that **MM-OPD** shifts evidence preference toward correct evidence while consistently narrowing the performance gap between image and symbolic conditioning.

5 Related Work

We briefly review the two lines of work most closely related to **MM-OPD**: on-policy self-distillation and distillation with privileged information. A more detailed discussion is provided in Appendix C.

On-Policy Self-Distillation. On-policy distillation provides dense supervision from a stronger teacher on the student's own rollouts, reducing the distribution mismatch of off-policy distillation (Agarwal et al., 2024; Lu & Lab, 2025; Li et al., 2026). Recent self-distillation methods obtain such supervision from the same model under richer conditioning (Zhao et al., 2026; Hübotter et al., 2026; He et al., 2026). While existing multimodal variants target regional perception or mathematical reasoning (Yuan et al., 2026; Yang et al., 2026), **MM-OPD** transfers symbolic-view guidance to image-conditioned reasoning to improve a broad range of multimodal abilities.

Distillation with Privileged Information. Privileged conditioning can strengthen self-teacher predictions but may also introduce preferences unavailable to the student. Existing methods either selectively route privileged supervision or remove preferences induced by privileged inputs (Yu et al., 2026b; Shen et al., 2026). **MM-OPD** instead uses captions and code as privileged symbolic views and applies their residual correction to image-conditioned predictions, improving evidence selection.

6 Conclusion

In this work, we revisit the assumption of a seamless transition from perception to reasoning. We identify the *Symbolic Visual Gap* and reveal *visual evidence selection* as one more bottleneck between perception and reasoning. Based on the finding that symbolic view elicits the model to assign answer-to-evidence attention to correct visual evidence, we introduce **MM-OPD**, a multimodal on-policy self-distillation framework for *symbolic-to-visual correction* that transfers guidance from symbolic views to the image-conditioned policy through residual token-level targets, steering it toward correct visual evidence. Experiments across benchmarks and model scales demonstrate improvements across a broad range of multimodal abilities rather than a single targeted capability.

AI USE STATEMENT

In this work, we did not use generative AI tools for any task with required disclosure. Specifically, generative AI was not used to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or assist in writing proofs, propose or refine hypotheses, design or provide feedback on our research methodology or experiments, implement methods, perform translation, clean or reformat datasets, support qualitative or thematic analysis, or interpret results; these uses are not applicable to this work. Among tasks with recommended disclosure, we used generative AI tools solely to aid and polish the writing, i.e., correcting grammar and spelling and improving the readability and phrasing of author-written text. We did not use generative AI for other recommended-disclosure tasks such as creating scientific figures, writing or editing code, searching for or identifying literature, formatting references, or brainstorming research ideas. We have reviewed all AI-assisted edits to ensure they are correct and faithfully preserve the authors' intended meaning. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4637–4646. IEEE, 2025.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Advances in Neural Information Processing Systems*, 38, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, and Guohai Xu. Deepeyesv2: Toward agentic multimodal model. In *International Conference on Learning Representations*, volume 2026, pp. 114851–114872, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

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. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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)*, pp. 20023–20070, 2025a.

Runqi Qiao, Qiuna Tan, Minghan Yang, Guanting Dong, Peiqing Yang, Shiqiang Lang, Enhui Wan, Xiaowan Wang, Yida Xu, Lan Yang, et al. V-thinker: Interactive thinking with images. *arXiv preprint arXiv:2511.04460*, 2025b.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, et al. Purified opsd: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*, 2025.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

XAI. Grok-1.5 vision preview, 2024. URL `https://x.ai/news/grok-1.5v`.

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Shilin Yan, Jintao Tong, Hongwei Xue, Xiaojun Tang, Yangyang Wang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. Act wisely: Cultivating meta-cognitive tool use in agentic multimodal models. *arXiv preprint arXiv:2604.08545*, 2026.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *Advances in Neural Information Processing Systems*, 38:29918–29952, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, et al. Dopd: Dual on-policy distillation. *arXiv preprint arXiv:2606.30626*, 2026b.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

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*, pp. 169–186. Springer, 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 48337–48383, 2025.

A Training Algorithm

Algorithm 1 **MM-OPD** Training

Input: Student π_θ, frozen self-teacher π_ϕ, training sample (I, Q, P), rollout number n, residual saturation scale γ

1: Sample n independent on-policy responses $\mathcal{Y} = \{y^{(j)}\}_{j=1}^{n}$ from $\pi_\theta(\cdot \mid I, Q)$ ▷ Eq. (1)
2: **for** each $y \in \mathcal{Y}$ and position t with prefix $h_t = y_{<t}$ **do**
3: Evaluate S_t, T_t, B_t, and R_t on h_t ▷ Eq. (2)
4: Select $\mathcal{K}_t \leftarrow \mathrm{Top}_K(T_t)$
5: Extract the centered privileged residual $\widetilde{d}_t$ ▷ Eq. (3)
6: Construct the residual-corrected distribution $\widehat{q}_t$ ▷ Eq. (4)
7: Preserve the anchor's head and tail masses to obtain $q_t^\star$ ▷ Eqs. (5) and (6)
8: Compress S_t onto $\mathcal{K}_t \cup \{\mathrm{tail}\}$ as $\widetilde{S}_t$
9: **end for**
10: Update θ by minimizing $\mathcal{L}_{\text{MM-OPD}}(\theta)$ ▷ Eq. (7)

B Details of Training Datasets Curation

To turn the symbolic visual gap into a training signal, we construct $\mathcal{D} = \{(I_i, Q_i, P_i)\}_{i=1}^{N}$, where P_i is privileged symbolic information, instantiated as a caption for a natural image or executable code for a structured visual. We use the following three-stage curation pipeline.

Cross-View Validation. We sample paired examples from V-Interaction (Qiao et al., 2025b), Mulberry (Yao et al., 2026), and ChartX (Xia et al., 2025), covering general VQA, mathematical reasoning, and chart understanding. The corresponding captions and code are taken directly from the source datasets. For natural images, we discard caption–image pairs whose contents are semantically inconsistent. For structured visuals, we execute the code and discard any sample whose code fails to execute, including those with syntax errors or missing dependencies.

Symbolic-View Normalization. The dataset-provided captions and code follow heterogeneous formats. We use Qwen3.5-397B to reorganize and normalize these existing symbolic views while preserving their original visual content; no new caption or code content is generated. Existing captions are reorganized into a consistent descriptive format, with boilerplate and image-irrelevant content removed. For code, we remove environment-specific elements, including hard-coded file paths and output operations such as `savefig`, while retaining the statements that construct the visual. This step reduces formatting variation and removes environment-dependent code across data sources.

Symbolic-View Utility Filtering. A valid image–symbolic pair does not necessarily improve the model's ability to answer its question. We estimate $\mathrm{Pass@8}$ with the original visual input (I, Q) and with the additional symbolic view (I, P, Q). We retain only examples satisfying $\mathrm{Pass@8}(I, Q) \leq \mathrm{Pass@8}(I, P, Q)$, excluding cases in which access to P lowers $\mathrm{Pass@8}$. The resulting dataset contains 32K examples spanning general VQA, mathematical diagrams, and charts.

C Additional Related Work

Multimodal Reinforcement Learning. Reinforcement learning with verifiable rewards (RLVR) (Guo et al., 2025) has become an important post-training approach for multimodal large language models (Hong et al., 2025). Recent work combines supervised cold-start training with RL or scales RL across diverse multimodal tasks (Huang et al., 2026; Tong et al., 2025). Agentic approaches further learn to acquire and manipulate visual evidence through cropping, zooming, and drawing (Hong et al., 2026; Su et al., 2026; Yan et al., 2026). However, final-answer rewards do not explicitly specify which visual evidence should support intermediate decisions. In contrast, **MM-OPD** uses token-level self-distillation from symbolic views to guide image-conditioned reasoning.

On-Policy Self-Distillation. On-policy distillation evaluates a stronger teacher on prefixes sampled by the current student, providing dense token-level supervision while reducing the distribution mismatch of off-policy distillation (Agarwal et al., 2024; Lu & Lab, 2025; Li et al., 2026). Self-

Distilled Reasoner (Zhao et al., 2026) conditions the teacher on reference solutions, SDPO (Hübotter et al., 2026) uses textual environment feedback, and Self-Distillation Zero (He et al., 2026) distills a self-reviser conditioned on the initial response and binary reward. In multimodal settings, Vision-OPD transfers crop-conditioned regional perception to a full-image student (Yuan et al., 2026), while RLSD combines self-distilled token-level update magnitudes with reward-based update directions (Yang et al., 2026). These methods primarily target specific capabilities, whereas **MM-OPD** uses symbolic-to-visual correction to improve diverse perception and reasoning tasks.

Distillation with Privileged Information. Privileged information can improve self-teacher predictions while introducing preferences based on inputs unavailable to the student. DOPD selectively routes supervision between privileged teacher and student policies according to token-level advantages and probabilities (Yu et al., 2026b). Purified OPSD discounts preferences induced by reference solutions and applies the remaining question-conditioned correction to a clean prediction (Shen et al., 2026). **MM-OPD** similarly controls information asymmetry, but uses captions and code as privileged symbolic views. Its residual target reduces preferences induced by the symbolic view alone and corrects a clean image-conditioned prediction toward better visual evidence selection.

D Implementation Details

Table 6 summarizes the training and inference configuration of **MM-OPD**; unless stated otherwise, these settings are used in all experiments. Controlled baselines reuse the same backbone, training data, and prompts; where applicable, we match the number of updates and the generated-token budget while keeping method-specific optimization settings. At inference, **MM-OPD** requires neither the symbolic view nor the self-teacher, decoding from the image and question alone.

Table 6: Training and inference configuration of **MM-OPD**.

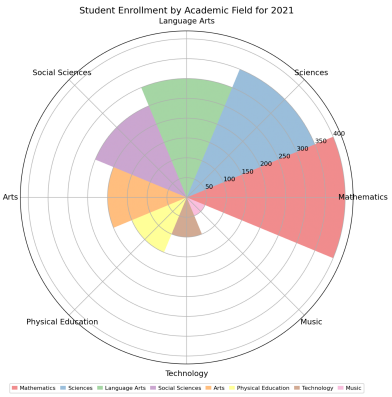
Configuration	Value
Backbones	Qwen3.5-4B, Qwen3.5-9B
Training set	MMOPD-32K
Self-teacher	Frozen
Distillation divergence	Jensen–Shannon (JSD)
Top-K support	$K = 100$
Residual saturation scale	$\gamma = 10$
On-policy rollouts per prompt	$n = 2$
Max response length	8192 tokens
Training epochs	1
Global batch size	256
Learning rate	1e$-$6
Hardware	8$\times$ H20 (141 GB)
Inference inputs	Image and question only

E Additional Case Studies

Figures 6 to 8 provide additional examples of visual evidence selection, supplementing Fig. 2 in the main text. Green and red mark correct and distracting evidence, respectively, with darker shading indicating stronger answer-to-evidence attention.

As shown in Fig. 6, the response contains the relevant chart values, yet associates a distracting value with the queried category and uses it in the derivation. The subsequent calculation is valid given that premise, but yields the wrong answer. Attention is also stronger on the distracting value used in the derivation. This case further distinguishes the availability of visual information from its selection as evidence.

A similar pattern appears in Fig. 7. The response repeatedly lists "Lack of control" across the three overlaps, but compares only pair-specific factors to conclude that there is no common cause. These distracting factors receive stronger attention, while the shared factor is omitted from the comparison. The required information is therefore already verbalized, but does not guide the final answer.



Question How many more students are enrolled in Mathematics compared to Social Sciences?

Ground truth 150 **Model answer** 200

Response

From the **radial bar chart**, we can estimate enrollment numbers based on the radial grid lines labeled **50**, **100**, **150**, **200**, **250**, **300**, **350**, **400**.

Step 1: Identify the values from the chart

- **Mathematics** (red segment) extends to the outermost ring, which corresponds to **400** students.
- **Social Sciences** (purple segment) extends to the ring labeled **200** students.

Step 2: Calculate the difference

Difference = Mathematics − Social Sciences = **400** − **200** = **200**

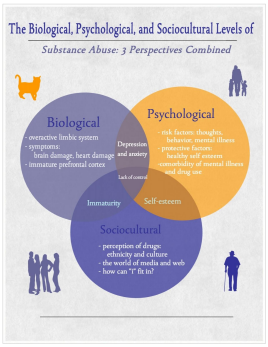
Final Answer: **200** more students are enrolled in Mathematics compared to Social Sciences.

answer-to-evidence attention: correct evidence distracting evidence

darker = stronger attention

Figure 6: Evidence selection in a radial bar chart. Relevant values are available, but distracting evidence is used as the premise for the derivation.

Question image



Question What emerges as the main cause of substance abuse from Biological, Sociocultural, and Psycological perspectives?

Ground truth Lack of control **Model answer** No common cause

Response

Step 1: Read the pairwise overlaps

- **Biological + Psychological:** **Depression and anxiety**; **Lack of control**.
- **Biological + Sociocultural:** **Immaturity**; **Lack of control**.
- **Psychological + Sociocultural:** **Self-esteem**; **Lack of control**.

Step 2: Identify the shared cause

Comparing the factors across the overlaps gives **Depression and anxiety**, **Immaturity**, and **Self-esteem**. Since each pair points to a different factor, there is no common cause.

Final answer: **No common cause**

answer-to-evidence attention: correct evidence corrupt evidence

darker = stronger attention

Figure 7: Evidence selection in a Venn diagram. The shared factor is verbalized but omitted from the final comparison.

Likewise, in Fig. 8, the response lists the required values 22, 25, and 39, yet selects 22 and 28 as the two smallest blue values and 55 as the smallest green value. The resulting comparison, $22+28 < 55$, yields "No" instead of the correct "Yes" from $22 + 25 > 39$. Attention also favors the distracting values used in the computation. Together, these examples support the observation in the main text: having task-relevant information available in the trace does not ensure that it is selected as evidence for subsequent reasoning.

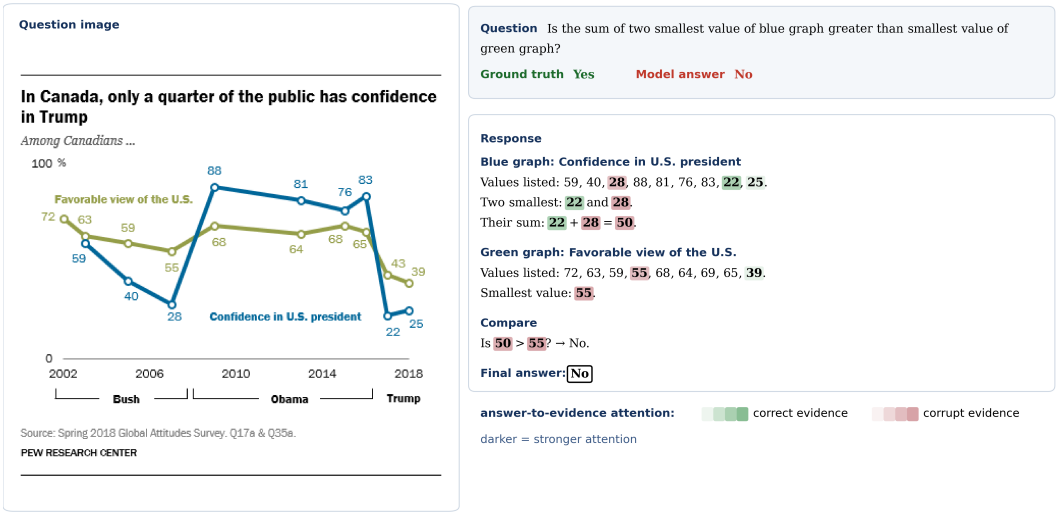


Figure 8: Evidence selection in a line chart. The required values are listed, but the computation uses incorrectly selected extrema.