Visual-Advantage On-Policy Distillation for Vision-Language Models

Ruiqi Liu[1,2] **Xiaolei Lv**[3] **Gengsheng Li**[1] **Ximo Zhu**[3] **Zhiheng Wang**[4]
Zhengbo Zhang[1] **Junkai Chen**[1] **Zhiheng Li**[1] **Bo Li**[3] **Jun Gao**[3] **Shu Wu**[1†]

[1]Institute of Automation, CAS [2]School of Advanced Interdisciplinary Sciences, UCAS
[3]Hello Group Inc. [4]Sun Yat-sen University

[†]Corresponding author: Shu Wu shu.wu@nlpr.ia.ac.cn

Abstract

On-policy knowledge distillation has proven effective for language models, yet its application to vision-language models (VLMs) remains underexplored. We observe that standard on-policy distillation can improve a student's output quality while producing only a small change in the teacher-scored visual sensitivity of its rollouts: task accuracy increases, yet the teacher's predictions along the student's realized tokens show little increase in their dependence on fine-grained visual detail. To make this difference observable, we introduce *visual advantage* (VA), the non-negatively rectified token-level log-probability difference when the teacher scores a student-generated rollout with the original versus pixelated image, keeping the question and preceding tokens fixed. VA is concentrated in a small minority of tokens, and controlled token-loss masking shows that these high-VA tokens carry a disproportionate share of useful visual supervision. This motivates a distillation objective that treats them differently from language scaffolding, so their contribution is not diluted by the abundant surrounding language tokens. We propose **Visual-Advantage On-Policy Distillation (VA-OPD)**, which uses VA at two granularities: rollout-level reweighting by trajectory-averaged VA, and token-level KL averaged within high-VA and low-VA groups separately. We train on two math datasets (Geometry3K and ViRL39K) and evaluate on eight benchmarks covering both mathematical reasoning and visual understanding, across three teacher sizes (4B, 8B, and 32B) on the Qwen3-VL family. VA-OPD improves over standard on-policy distillation on every benchmark, with aggregate gains increasing along the tested teacher-size and data-scale comparisons. Independent OCR-token and attention analyses further support the association between VA and image-conditioned generation. Code is available at GitHub.

1 Introduction

On-policy distillation has emerged as an effective strategy for compressing large language models [1–8]: a student learns by matching a teacher's output distribution on its own rollouts. Extending it to vision-language models (VLMs) [9], however, surfaces a mismatch that text-only distillation never had to confront. Not all tokens in a VLM's response are created equal with respect to the image. As illustrated in Figure 1 (left), teacher-scored sensitivity to fine-grained visual detail is concentrated on a small subset of positions in the analyzed rollouts, while the remainder includes language scaffolding, reasoning templates, and information that survives the degradation. Standard KL-based distillation weights every token uniformly, raising a basic question: does on-policy distillation actually push the student to rely on visual input, or is matching the teacher on the language scaffolding alone enough?

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

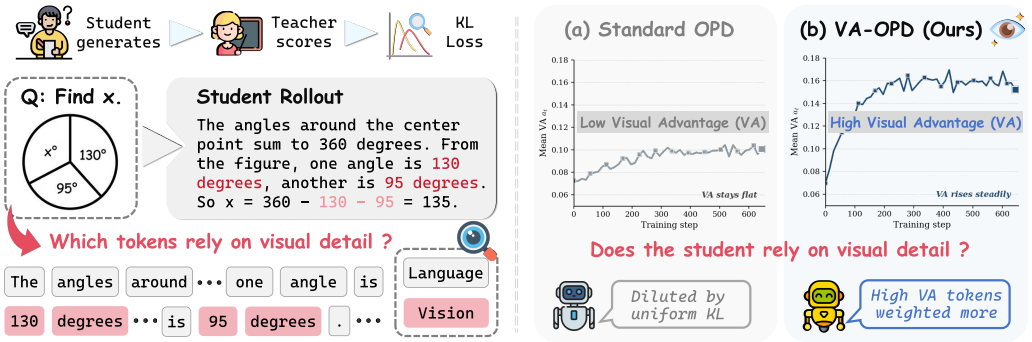


Figure 1: **Token-level visual reliance in VLM distillation.** *Left:* in a student rollout, a minority of positions (red) receive high teacher-scored VA, while the remainder (gray) includes language-template tokens. Standard OPD applies a uniform KL weight to all tokens, diluting the learning signal on vision-critical positions. *Right:* teacher-scored visual advantage (VA), averaged over the student's rollouts, tracked across training, for two distillation methods. Under **Standard OPD**, student-rollout VA stays nearly flat: matching the teacher's outputs yields only a small change in this visual-sensitivity diagnostic on student-generated rollouts. Under **VA-OPD**, student-rollout VA increases during training, consistent with stronger sensitivity to visual detail at generation positions. Independent validation is reported in Section 5.

Answering this requires a token-level signal of visual dependency that a uniform KL cannot supply on its own. We introduce a simple counterfactual probe: for each token in a student-generated rollout, we score it under the teacher with and without fine-grained visual detail, and take the non-negatively rectified log-probability difference as the per-token *visual advantage* (VA). VA quantifies, for each student-generated token, the positive change in the teacher's conditional log-probability between the original and pixelated images, with the question and realized prefix fixed. Tracking VA on student rollouts during training reveals a concerning pattern (Figure 1, right): standard on-policy distillation steadily improves task accuracy, yet rollout VA stays nearly flat. Task accuracy improves under teacher supervision, but teacher-scored sensitivity to fine-grained visual detail changes little. Evaluating these paired teacher predictions along the student's realized prefixes provides a consistent diagnostic throughout training.

Two observations on student rollouts help explain this. VA is heavy-tailed: the top 10% of tokens carry most of the total VA mass while the rest sit close to zero. A controlled mask ablation shows that this small minority concentrates useful visual supervision, since removing the KL loss on the top-VA tokens substantially lowers accuracy while removing the same fraction of low-VA or random tokens has a much smaller effect. Standard KL averages uniformly over all tokens, so the loss weight assigned to the few high-VA tokens is diluted by the surrounding linguistic scaffolding.

We propose **Visual-Advantage On-Policy Distillation (VA-OPD)**, which uses VA at two levels of granularity to address this dilution. At the rollout level, the K sibling rollouts under the same prompt vary in how much their tokens depend on fine-grained visual detail; we soft-weight each rollout by its trajectory-averaged VA relative to its siblings, so that rollouts with stronger teacher-scored visual sensitivity receive higher distillation loss weights. At the token level, we sort each rollout's tokens by VA, assign the top fraction and remainder to high- and low-VA groups, then separately average per-token reverse KL per group before summing. Both groups retain the teacher distribution conditioned on the original image as their target; the pixelated image supplies only the VA weighting signal. Beyond Standard OPD, VA-OPD only adds one teacher forward pass on the image with fine-grained visual detail suppressed; it needs no extra training annotations, reward models, or student rollouts.

We validate VA-OPD across three teacher sizes (4B, 8B, and 32B), two training corpora (Geometry3K and ViRL39K), and eight benchmarks. VA-OPD improves over standard on-policy distillation on every benchmark. Gains are sizable throughout the math-reasoning benchmarks, and within visual understanding the largest gains are HallusionBench $+2.5$ and AI2D $+2.4$, benchmarks probing image-grounded hallucination and diagram reasoning. The gain amplifies under the larger ViRL39K corpus (Math Avg $+3.8$, Visual Avg $+2.5$). During VA-OPD training, VA on student rollouts rises alongside accuracy, consistent with stronger visual sensitivity at generation positions alongside better performance on the evaluated tasks.

Our contributions are:

- We find that teacher-scored VA is concentrated on a small minority of positions in the analyzed VLM rollouts, and uniform-weighted KL gives this subset the same per-token loss coefficient as the surrounding positions. In our controlled comparison, task accuracy improves under teacher supervision while teacher-scored VA on student rollouts changes little.
- We introduce *visual advantage* (VA), which characterizes teacher-relative visual sensitivity at each token in student-generated rollouts. Based on this, we propose VA-OPD, which applies VA at two granularities: rollout-level reweighting by trajectory-averaged VA, and token-level KL averaged within high-VA and low-VA groups separately.
- Across three teacher sizes (4B, 8B, and 32B), two training corpora (Geometry3K and ViRL39K), and eight benchmarks, VA-OPD improves over standard on-policy distillation on every benchmark. The teacher-scored VA on student rollouts also rises with accuracy during VA-OPD training, showing that the gain is accompanied by stronger teacher-scored visual sensitivity.

2 Background

We consider on-policy distillation of vision-language models: a frozen teacher p_{T} and student p_{S}, both conditioned on an image v and query q. The student generates K rollouts $\{y^{(k)}\}_{k=1}^{K}$ with the decoding settings of Appendix A.1; KL supervision uses the models' full next-token distributions. The teacher provides token-level targets via the reverse KL [2, 10]:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}(p_{\mathrm{S}}(\cdot \mid v, q, y_{<t}) \,\|\, p_{\mathrm{T}}(\cdot \mid v, q, y_{<t}))\,. \tag{1}$$

We refer to training with this objective throughout the paper as *Standard OPD*. The objective assigns the same coefficient to every response position, including language-template tokens and positions that depend on fine-grained visual detail. Matching the surrounding language scaffolding can therefore reduce the average KL while providing limited emphasis on information in the diagram. We now introduce a diagnostic that makes teacher-relative visual sensitivity observable (Section 2.1), then present empirical evidence that motivates our method (Section 2.2).

2.1 Visual Advantage

For each token in a student rollout, we want to know how the teacher's conditional prediction changes when fine-grained visual detail is removed. The probe is simple: score the same token under the teacher in two matched conditions, with and without fine-grained visual detail, and take the non-negatively rectified log-probability difference as the per-token *visual advantage* (VA):

$$a_t = \max(\log p_{\mathrm{T}}(y_t \mid v, q, y_{<t}) - \log p_{\mathrm{T}}(y_t \mid \tilde{v}, q, y_{<t}),\ 0)\,, \tag{2}$$

where $\tilde{v}$ is a degraded version of v with fine-grained visual detail suppressed. A high a_t means that the original image raises the teacher's conditional probability of the realized token, e.g., a precise number that must be read off the diagram; $a_t \approx 0$ means little positive change under this probe. Such positions can involve language scaffolding or robust global image information that survives degradation. VA therefore characterizes the teacher's conditional sensitivity to the visual detail removed by the probe at each student-generated position. Zero rectification retains positive original-image effects and assigns zero VA to positions where the degraded view gives the token a higher log-probability.

We realize $\tilde{v}$ by downsampling each image dimension to 10% with bilinear interpolation and upsampling back via nearest-neighbor. This suppresses fine-grained visual detail (numerical labels, OCR, small symbols) while preserving broad layout and color, and keeps the visual token count unchanged so that the logprobs under v and $\tilde{v}$ align token-for-token. The broad layout and color retained after degradation preserve coarse visual context. Matching dimensions and token counts controls input-format differences between the two conditions, yielding a structurally aligned probe of fine-grained information. The construction comparisons in Appendix B assess this choice empirically.

Function-word examples are discussed in Appendix C.

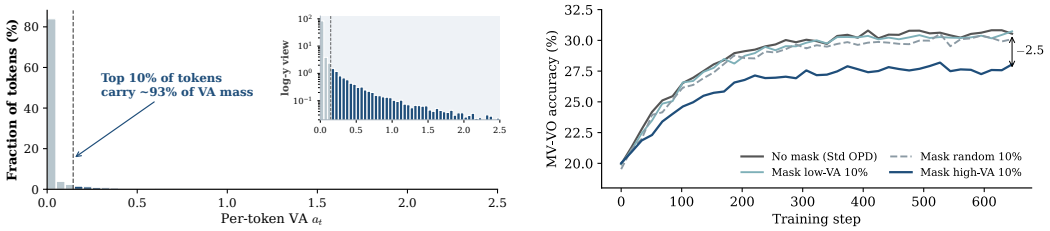


Figure 2: **Motivating observations on student-generated rollouts. (a)** VA is heavy-tailed: the top 10% of tokens (right of the dashed line) carry $\sim$93% of total VA mass. **(b)** MathVerse Vision Only training curves for Standard OPD and three 10%-token-masking variants; high-VA masking substantially drops accuracy, while low-VA or random masking has little effect.

2.2 Motivating Observations

Key insight. Visual supervision in the analyzed VLM rollouts is concentrated in a small minority of tokens, and these tokens carry a signal relevant to learning image-dependent predictions. Masking their KL loss substantially degrades performance, while masking an equal fraction of low-VA or random tokens has a much smaller effect. Standard on-policy distillation weights all tokens uniformly and therefore dilutes this concentrated signal with surrounding language scaffolding.

We establish two empirical observations that motivate the token-level design of VA-OPD, with both measured directly on student-generated rollouts to reflect the student's own generation behavior.

Observation 1: VA is sparse within a rollout. For each student rollout, we compute a_t at every position and sort tokens by VA. Figure 2(a) shows the resulting distribution, aggregated across $\sim$105k tokens of held-out rollouts from a Standard OPD student: the top 10% of tokens carry $\sim$93% of the total VA mass, while the rest have VA close to zero. This long-tail structure indicates that teacher-scored visual sensitivity is sparse on these VLM rollouts.

Observation 2: High-VA tokens carry the visual supervision signal. We design a controlled mask ablation: three Standard OPD variants (8B$\rightarrow$2B on Geometry3K) that differ only in which 10% of tokens have their per-token KL *masked out*: (i) a random 10%, (ii) the lowest-VA 10%, (iii) the highest-VA 10%. All other training conditions are identical. Figure 2(b) shows MathVerse Vision Only training curves. At the displayed endpoints (approximately 650 steps), low-VA masking is about 0.2 points above the no-mask control, random masking is about 0.4 points below it, and masking the highest-VA 10% is about 2.5 points below it, despite excluding only 10% of response positions. The high-VA minority therefore carries not just a disproportionate share of VA mass within each rollout, but also supervision whose removal most strongly affects visual-task accuracy in this controlled comparison.

Together, these observations help explain why standard on-policy distillation can improve accuracy with only a small change in teacher-scored visual sensitivity: the visual supervision signal is concentrated in a small token minority, and under a uniform KL mean every per-token KL term receives coefficient $1/T$, so the concentrated signal is diluted by the surrounding scaffolding. We address this in Section 3 by sorting tokens by VA and normalizing KL within high-VA and low-VA groups separately.

3 Visual-Advantage On-Policy Distillation

We propose **Visual-Advantage On-Policy Distillation (VA-OPD)** (Figure 3), which uses VA at two granularities. At the rollout level (Section 3.1), each rollout is soft-weighted by its trajectory-averaged VA relative to its sibling rollouts under the same prompt, so that rollouts with higher teacher-scored mean VA receive larger distillation loss weights. At the token level (Section 3.2), tokens within each rollout are sorted by VA into a high-VA group V and a low-VA group L, and the per-token KL is averaged within each group separately before being summed. The full objective is given in Section 3.3.

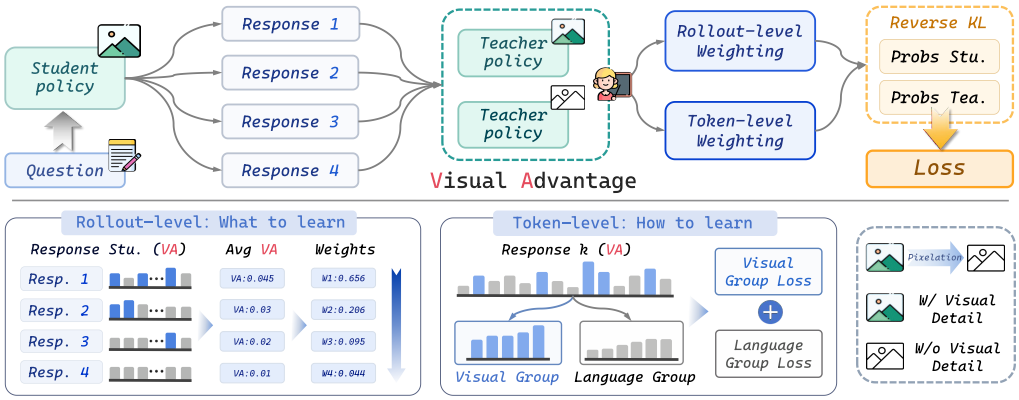


Figure 3: **Overview of Visual-Advantage On-Policy Distillation (VA-OPD).** The teacher scores each token with and without fine-grained visual detail; the log-prob difference gives the per-token visual advantage (VA). VA drives reverse-KL distillation at two granularities: rollout-level reweighting by relative VA (*what to learn*) and token-level KL split into high- and low-VA groups (*how to learn*). Both groups use the original-image teacher as the KL target; the degraded view supplies only the VA signal.

3.1 Rollout-Level Reweighting

Even when sampled from the same student under fixed (v, q), sibling rollouts can diverge in their token sequences and in the teacher-scored sensitivity of those positions to fine-grained visual detail; their trajectory-averaged VA $\bar{a}^{(k)}$ can therefore vary within a K-rollout group. We exploit this within-group spread by assigning each rollout a soft weight derived from its visual advantage relative to its $K-1$ siblings, so that more loss weight flows to rollouts with higher teacher-scored mean VA.

For each rollout $y^{(k)}$, we compute the trajectory-averaged VA and normalize it within the K sibling rollouts:

$$\bar{a}^{(k)} = \frac{1}{T^{(k)}} \sum_{t=1}^{T^{(k)}} a_t^{(k)}, \qquad \hat{z}^{(k)} = \frac{\bar{a}^{(k)} - \mu}{\sigma + \epsilon}, \tag{3}$$

where μ and σ are the mean and population standard deviation (denominator K) of $\{\bar{a}^{(j)}\}_{j=1}^{K}$, and $\epsilon = 10^{-6}$ is a small constant for numerical stability. A positive $\hat{z}^{(k)}$ indicates higher teacher-scored mean VA than the siblings' mean within the same candidate set; a negative value indicates relatively lower mean VA. We then convert $\hat{z}^{(k)}$ into rollout-level distillation weights via a temperature-controlled softmax with $\tau = 1.0$:

$$w^{(k)} = \frac{\exp\big(\hat{z}^{(k)}/\tau\big)}{\sum_{j=1}^{K} \exp\big(\hat{z}^{(j)}/\tau\big)}, \qquad \sum_{k=1}^{K} w^{(k)} = 1. \tag{4}$$

3.2 Token-Level Grouped KL

Within the analyzed rollouts, high-VA positions form a small minority; many surrounding tokens provide language scaffolding or information less sensitive to the fine detail removed by our probe. Under a uniform KL mean over all T tokens, every per-token KL term receives coefficient $1/T$, so adding surrounding scaffolding reduces the loss weight assigned to a fixed set of vision-dependent positions. We address this by sorting each rollout's tokens by VA, treating the top fraction as a high-VA group and the rest as a low-VA group, and averaging per-token KL within each group separately.

For each rollout $y^{(k)}$ of length $T^{(k)}$, we rank its tokens by VA $\{a_t^{(k)}\}_{t=1}^{T^{(k)}}$ and place the top-p_v fraction in the high-VA group $V^{(k)}$; the remaining tokens form the low-VA group $L^{(k)}$. The hyperparameter $p_v \in (0, 1)$ controls the relative size of V and we fix $p_v = 0.2$ for the primary

experiments. For nonempty high- and low-VA groups, the per-rollout loss is a weighted sum of the two within-group KL averages:

$$\mathcal{L}_{\text{group}}^{(k)} = \lambda \cdot \frac{1}{|V^{(k)}|} \sum_{t \in V^{(k)}} \mathrm{KL}_t \;+\; (1-\lambda) \cdot \frac{1}{|L^{(k)}|} \sum_{t \in L^{(k)}} \mathrm{KL}_t, \tag{5}$$

where $\mathrm{KL}_t = \mathrm{KL}(p_{\mathrm{S}}(\cdot \mid v, q, y_{<t}) \,\|\, p_{\mathrm{T}}(\cdot \mid v, q, y_{<t}))$ is the per-token reverse KL from Equation (1), and $\lambda \in (0, 1)$ controls the high-VA group's weight. We use $\lambda = 0.5$ as the default. Since each group is size-normalized, its total loss coefficient is λ or $1 - \lambda$, independently of the other group's size. The resulting per-token loss coefficients are $\lambda/|V|$ in V and $(1 - \lambda)/|L|$ in L, preserving each group's assigned loss mass as its size changes. The degraded image $\tilde{v}$ is used only for VA computation; the KL target in Equation (5) always uses the teacher distribution conditioned on the original image v.

The default weighting and grouping choices are further discussed in Appendix A.7.

3.3 Overall Objective

Combining rollout-level weights with token-level grouped KL gives the per-prompt VA-OPD loss for $x = (v, q)$:

$$\mathcal{L}_{\text{VA-OPD}}(x) = \sum_{k=1}^{K} w^{(k)} \left[\lambda \cdot \frac{1}{|V^{(k)}|} \sum_{t \in V^{(k)}} \mathrm{KL}_t + (1-\lambda) \cdot \frac{1}{|L^{(k)}|} \sum_{t \in L^{(k)}} \mathrm{KL}_t \right], \tag{6}$$

with full training objective $\mathcal{L}_{\text{VA-OPD}} = \mathbb{E}_{x \sim \mathcal{D}} \left[\mathcal{L}_{\text{VA-OPD}}(x) \right]$. VA-OPD needs no additional training annotations, reward models, or student rollout changes; the teacher performs only one extra forward pass per rollout without fine-grained visual detail to supply VA signals, leaving the inference procedure unaffected. The update convention is described in Appendix A.6.

4 Experiments

We ask: (Q1) does VA-OPD beat Standard OPD and RL baselines on accuracy (Section 4.2); (Q2) does it generalize across scales and data (Section 4.3); (Q3) which components drive the gain (Section 4.4); and (Q4) is VA-OPD's added per-step compute justified by its accuracy gain (Section 4.5)?

4.1 Experimental Setup

Models and data. Our primary setting uses Qwen3-VL-8B-Instruct [11] as the teacher and Qwen3-VL-2B-Instruct as the student, trained on Geometry3K [12] ($\sim$2.1K problems). We further test generalization across teacher sizes (Qwen3-VL-4B$\rightarrow$2B and Qwen3-VL-32B$\rightarrow$2B) and on a larger dataset ViRL39K [13] ($\sim$39K problems). The counterfactual teacher distribution $\tilde{p}_{\mathrm{T}}$ used to compute VA follows the construction in Section 2.1.

Training. We train for 5 epochs using AdamW with $K = 4$ rollouts per prompt and batch size 16. VA-OPD defaults are $\tau = 1.0$ (rollout-reweighting softmax temperature), $\lambda = 0.5$ (high-VA group weight), $p_v = 0.2$ (high-VA fraction), and pixelation ratio 0.10; these are fixed across all configurations with no per-configuration tuning, and trained distillation methods within a configuration use matched initialization and optimization settings, the same underlying training corpus, and identical within-benchmark evaluation protocols, facilitating comparison of the training objectives. Training, hyperparameter, and compute-comparison details are reported in Appendix A.

Baselines and benchmarks. We compare VA-OPD against six baselines evaluated under matched settings: Base (the un-distilled student evaluated directly), CoT-SFT (Appendix A.4), Off-policy KD [14], Standard OPD [2] (Equation (1)), GRPO [15], and PAPO [16]. We evaluate on eight benchmarks grouped into mathematical reasoning (WeMath [17], MathVista [18], MathVerse [19]) and visual understanding (HallusionBench [20], AI2D [21], MMMU [22], MMStar [23], OCRBench [24]). We report avg@8 at temperature 1.0 under official scoring protocols using VLMEvalKit [25], with GPT-4o judging MathVista, MathVerse, and HallusionBench (Appendix A.2).

Table 1: **Main results on eight benchmarks.** Primary setting: Qwen3-VL-8B→Qwen3-VL-2B, trained on Geometry3K. Checkpoints are selected on 200 held-out Geometry3K problems. Scores are avg@8 (temperature 1.0); per-benchmark metric conventions follow each benchmark's official protocol (Appendix A.2). MV-VO denotes `MathVerse_MINI_Vision_Only` (788 inputs). The highest score in each column is bold.

Method	Math Reasoning				Visual Understanding					
	WeMath	MathVista	**MV-VO**	**Avg**	HalluB	AI2D	MMMU	MMStar	OCRBench	**Avg**
Base	36.8	63.9	19.6	40.1	52.6	77.8	49.1	56.1	85.9	64.3
CoT-SFT	39.7	64.4	23.9	42.7	51.0	75.7	48.7	57.1	86.2	63.7
Off-policy KD	38.9	65.3	24.3	42.8	52.3	76.0	49.3	57.4	85.5	64.1
GRPO	44.1	64.6	24.9	44.5	54.0	77.5	**52.9**	57.8	84.8	65.4
PAPO	44.8	64.5	25.8	45.0	54.0	77.8	52.7	58.0	84.7	65.4
OPD	43.3	63.7	29.1	45.4	52.0	75.8	50.9	59.7	84.7	64.6
VA-OPD (ours)	**46.6**	**66.4**	**31.9**	**48.3** ↑2.9	**54.5**	**78.2**	51.5	**59.9**	**86.4**	**66.1** ↑1.5

Table 2: **Standard OPD vs. VA-OPD across four configurations** (avg@8). All configurations use Qwen3-VL and the matched evaluation setting of Table 1; MV-VO denotes `MathVerse_MINI_Vision_Only`. Metric conventions follow Appendix A.2.

T→S	Data	Method	Math Reasoning				Visual Understanding					
			WeMath	MathVista	**MV-VO**	**Avg**	HalluB	AI2D	MMMU	MMStar	OCRBench	**Avg**
4B→2B	Geo3K	OPD	43.5	65.4	26.9	45.3	52.5	75.0	49.2	60.4	84.8	64.4
		VA-OPD	**45.9**	**67.3**	**29.0**	**47.4** ↑2.1	**53.7**	**75.8**	**49.9**	**61.1**	**85.5**	**65.2** ↑0.8
8B→2B	Geo3K	OPD	43.3	63.7	29.1	45.4	52.0	75.8	50.9	59.7	84.7	64.6
		VA-OPD	**46.6**	**66.4**	**31.9**	**48.3** ↑2.9	**54.5**	**78.2**	**51.5**	**59.9**	**86.4**	**66.1** ↑1.5
32B→2B	Geo3K	OPD	54.1	67.6	31.5	51.1	52.3	77.0	50.0	62.5	84.7	65.3
		VA-OPD	**57.5**	**70.8**	**36.1**	**54.8** ↑3.7	**55.1**	**79.5**	**51.0**	**64.0**	**86.9**	**67.3** ↑2.0
8B→2B	ViRL39K	OPD	46.1	63.3	29.7	46.4	53.6	75.4	51.4	61.8	85.2	65.5
		VA-OPD	**50.3**	**67.5**	**32.7**	**50.2** ↑3.8	**55.8**	**79.5**	**53.2**	**64.0**	**87.5**	**68.0** ↑2.5

4.2 Main Results

Table 1 reports the primary comparison of all seven methods across eight benchmarks (8B→2B, trained on Geometry3K only), with model scale and the underlying training-question pool held fixed.

VA-OPD consistently outperforms Standard OPD across all benchmarks. Compared to Standard OPD, the strongest distillation baseline on the aggregate scores, VA-OPD improves on all eight benchmarks spanning both mathematical reasoning and visual-understanding tasks: WeMath $+3.3$, MathVista $+2.7$, MV-VO $+2.8$, HallusionBench $+2.5$, AI2D $+2.4$, MMMU $+0.6$, MMStar $+0.2$, and OCRBench $+1.7$. The aggregate effect is $+2.9$ on Math Avg and $+1.5$ on Visual Avg, showing that the advantage is broad rather than driven by one outlier; within visual understanding, gains concentrate on HallusionBench and AI2D, where answers hinge on careful image reading, a pattern consistent with stronger visual dependency during distillation improving image-grounded answers on hallucination and diagram-understanding tasks.

The headline gain pattern tracks visual dependency. Among the visual-understanding benchmarks, the largest gains occur on HallusionBench (image-grounded hallucination probing) and AI2D (diagram-based reasoning), where accurate answers depend on reading visual entities and relations. MMMU and MMStar see noticeably smaller gains, suggesting that the benefit can vary with the visual information required and the student's existing capabilities. This pattern is consistent with VA-OPD's mechanism: by sorting tokens by teacher-scored VA, the method concentrates the distillation signal on positions whose predictions are sensitive to fine-grained visual detail, while retaining supervision on the low-VA language scaffolding. The independent attention, OCR-token, and perturbation analyses in Appendix C further examine this visual-dependency interpretation.

4.3 Generalization across Scales and Data

Table 2 extends the comparison to four configurations that vary training-data scale and teacher size (with the student fixed at 2B).

Larger training corpus does not dilute the gain. Holding the 8B→2B teacher–student pair fixed and switching from Geometry3K ($\sim$2.1K problems) to the substantially larger ViRL39K ($\sim$39K

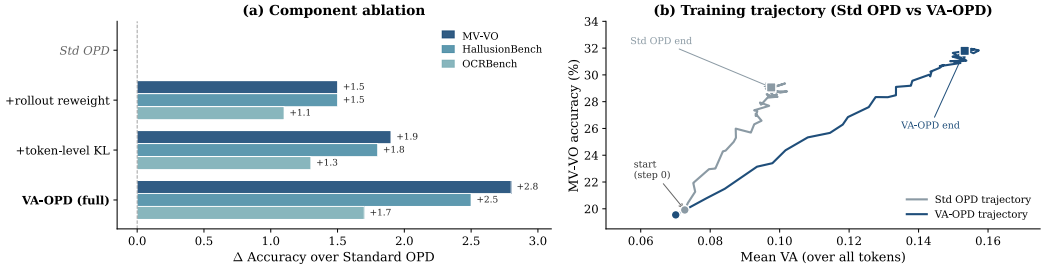


Figure 4: **Ablation and training-trajectory analysis.** (a) Per-benchmark Δ over Standard OPD for two VA-OPD components alone and combined, separating their individual and joint contributions: rollout-level reweighting only, token-level grouped KL only, and full VA-OPD. (b) Training trajectories in $(\text{mean VA over all tokens}, \text{MV-VO accuracy})$ space throughout training for Standard OPD vs. VA-OPD students. Circles mark the start (initial Qwen3-VL-2B-Instruct), squares the end of training. The trajectories move in distinctly different directions: Standard OPD climbs nearly vertically (accuracy rises while mean VA stays near the start), whereas VA-OPD moves diagonally up-right (accuracy and mean VA rise together, indicating coupled improvement).

problems), changing both the corpus scale and its example composition, VA-OPD's Math Avg gain over Standard OPD grows from $+2.9$ to $+3.8$ while Visual Avg grows from $+1.5$ to $+2.5$. With roughly $20\times$ as much training data and broader example coverage, VA-OPD achieves larger gains on both aggregates under the same teacher–student pair. This is consistent with broader visual coverage offering additional visually dependent positions for VA-OPD to exploit.

Gain holds across teacher sizes. Holding the student fixed at 2B and varying the teacher across 4B, 8B, and 32B on Geometry3K, thereby isolating teacher capacity as the varying factor, the gain of VA-OPD over Standard OPD stays clearly positive at all three scales: $+2.1/+0.8$ at 4B, $+2.9/+1.5$ at 8B, and $+3.7/+2.0$ at 32B for Math/Visual Avg, a pattern consistent with stronger teachers supplying richer counterfactual signals for VA-OPD to extract. This shows that the observed effect is not scale-specific. Complete MathVerse evaluation and additional student/model-family tests are reported in Appendix A.5.

4.4 Ablation Study and Controls

We validate VA-OPD along two complementary axes (Figure 4): (a) a component ablation on MV-VO (vision-intensive math), HallusionBench (hallucination), and OCRBench (OCR) that isolates each VA-OPD component and its contribution to overall gains; (b) a training trajectory comparison that plots each student's path through $(\text{VA}, \text{accuracy})$ space across training, contrasting Standard OPD with VA-OPD and revealing whether accuracy and VA co-evolve.

Both components are needed for the full gain. Removing either VA-OPD component (rollout reweighting or token-level grouped KL) measurably degrades performance, showing that neither mechanism alone recovers the full measured gain; only their joint combination delivers the full VA-OPD gain. Each component alone nevertheless recovers a sizable fraction of the improvement, consistent with the two contributing through nontrivial and largely complementary pathways. Full VA-OPD reaches $+2.8$, $+2.5$, and $+1.7$ on MV-VO, HallusionBench, and OCRBench respectively, peaking on MV-VO where solving the problem demands reading off precise diagram quantities directly from the image, a setting in which visual grounding is particularly relevant.

VA and accuracy co-evolve during training. Figure 4(b) plots each student's path in $(\text{mean VA}, \text{MV-VO accuracy})$ space, where mean VA is averaged over *all* tokens rather than selected high-VA positions. Standard OPD traces a nearly *vertical* path: accuracy rises from 19.6 to 29.1 while the all-token mean VA only inches up ($0.07 \rightarrow 0.10$). VA-OPD instead traces a *diagonal* path: accuracy climbs higher (to 31.9) *and* the all-token mean VA grows substantially ($0.07 \rightarrow \sim 0.15$). The two trajectory shapes make the different training behavior visible under this all-token diagnostic: Standard OPD gains accuracy with a relatively small increase in mean VA, whereas VA-OPD couples a larger accuracy gain with a substantially larger VA increase across the rollout. Independent analyses complement this trend (Appendix C).

4.5 Training Efficiency

Figure 5 plots MV-VO accuracy against wall-clock time on the primary 8B→2B Geo3K configuration. The two curves are similar for roughly the first ∼3 h, after which VA-OPD remains above Standard OPD, so the resulting gain manifests in two complementary ways: (i) at a fixed accuracy target, VA-OPD matches Standard OPD's final accuracy of 29.1 in only ∼6.5 h versus 19.3 h, yielding a ∼3.0× wall-clock speedup *to that accuracy target*; (ii) VA-OPD reaches a higher final ceiling of 31.9, fully +2.8 points above Standard OPD's endpoint. The early-stage similarity suggests that the visual-dependency signal may become more useful as the student first learns the teacher's surface patterns. In this interpretation, high-VA tokens can become a bottleneck for further gains, and VA-OPD's redistribution of distillation loss weight toward these positions can help sustain continued improvement.

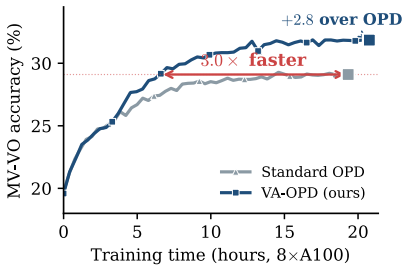


Figure 5: **MV-VO accuracy vs. training time on 8×A100 (Qwen3-VL-8B→2B, Geo3K).** The arrow marks time to Standard OPD's final accuracy.

5 Qualitative Token-Level VA Visualization

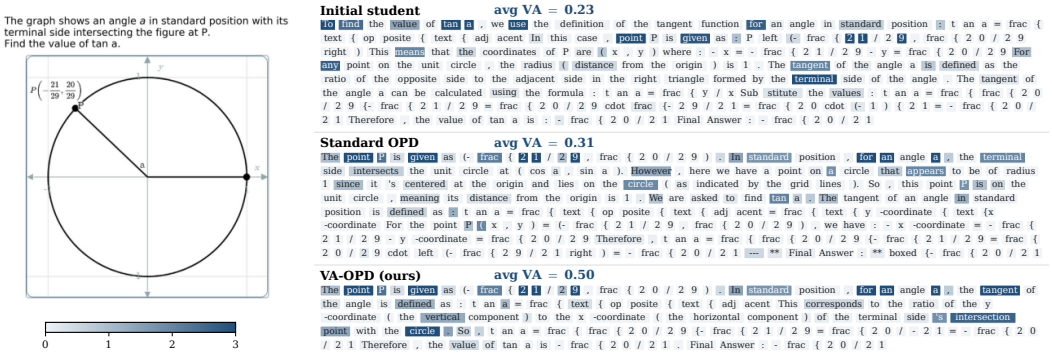


Figure 6: **Token-level VA visualization** on a representative geometry input. Each token is colored by its teacher-scored VA in responses from three students at different stages: the initial student (Base, before distillation), the Standard OPD student, and the VA-OPD student (darker = higher VA_t). All three responses contain high-VA positions; their displayed rollout means are 0.23, 0.31, and 0.50, respectively. The colors show how fine-grained image detail affects the teacher's predictions at the generation positions in each response.

Figure 6 inspects the visual advantage at the token level on a held-out geometry input. The initial student (before any distillation) and Standard OPD already exhibit VA hotspots at some numerical and phrase positions. In this example, the displayed rollout-averaged VA is 0.23 for Base, 0.31 for Standard OPD, and 0.50 for VA-OPD; the location and strength of hotspots vary across the generated responses. Concretely, the task requires reading the point coordinates $P = (-21/29, 20/29)$ from the diagram, identifying the terminal side in the second quadrant, and computing the coordinate ratio $\tan a = y/x = -20/21$. Highlighted positions include coordinate values and descriptions of the point and terminal side, while many surrounding language-template positions remain pale; function-word exceptions are discussed in Appendix C. This per-token contrast complements the controlled mask experiment in Section 2.2: the mask experiment showed that a small VA-aware token subset carries a disproportionate share of supervision useful for visual-task accuracy, and the heatmap illustrates teacher-scored sensitivity at numerical positions and their surrounding geometric explanation. Together, the quantitative ablation and qualitative heatmap support the same design rationale: VA-OPD increases the relative loss weight of positions whose teacher-scored probabilities are sensitive to fine-grained visual detail. Each response is scored under its own realized prefix, so the heatmap follows the token sequence generated by the corresponding student.

Independent OCR-token annotations and attention analyses complement this visualization; their protocols, complete results, and behavioral controls are reported in Appendix C.

6 Related Work

Visual dependency in VLMs. VLMs often fall back on language priors instead of attending to fine-grained visual detail [9, 26–30]. Mitigations include inference-time contrastive decoding [26–28, 31], preference-based fine-tuning [29, 32], and RL-based methods that inject perception into the policy gradient [16, 33–35]. Based on this, we propose VA-OPD, which applies VA at two complementary granularities: rollout-level reweighting by trajectory-averaged VA, and token-level KL averaged separately within high-VA and low-VA groups.

On-policy distillation for LLMs. GKD [1] and MiniLLM [2] established on-policy distillation via teacher matching on student rollouts. Subsequent work refines the supervision [3, 7, 8, 10, 36–41], the sampling strategy [5, 6, 42, 43], or identifies important tokens by teacher penalty or entropy [44, 45]. These studies focus on text-only models; those using token importance rely on teacher-derived or entropy-based signals rather than an image-conditioned counterfactual; by contrast, VA-OPD extends this paradigm to VLMs and shows that token-uniform objectives miss the sparse concentration of visual supervision within visually grounded VLM rollouts.

Positioning. VA-OPD sits at the intersection of these two lines, and to our knowledge is the first to incorporate per-token teacher-scored reliance on fine-grained visual detail into a two-level on-policy distillation objective for VLMs. Unlike inference-time methods, it modifies the weighting and grouping of the training loss while preserving the original-image teacher target, with no test-time overhead; unlike RL-based perception methods, it leverages teacher supervision rather than verifiable rewards. Differences from VPPO are detailed in Appendix D.

7 Conclusion

We showed that standard on-policy distillation for VLMs can improve task accuracy while producing only a small increase in teacher-scored visual sensitivity on student rollouts: the visual supervision concentrated on a small minority of tokens is diluted by uniform-weighted KL averaging across the surrounding language scaffolding. To observe this insensitivity, we introduced visual advantage (VA), a teacher-relative counterfactual measuring token-level dependency on fine-grained visual detail. Building on VA, we proposed Visual-Advantage On-Policy Distillation (VA-OPD), which uses VA at two granularities: rollout-level reweighting that assigns larger distillation loss weights to rollouts with higher teacher-scored sensitivity to fine-grained visual detail, and token-level KL averaged separately within high-VA and low-VA groups so the concentrated visual signal is not diluted by the surrounding scaffolding. Across three teacher sizes (4B, 8B, and 32B), two training corpora (Geometry3K and ViRL39K), and eight benchmarks, VA-OPD consistently improves over standard on-policy distillation, with aggregate gains increasing in the tested teacher-size and training-corpus comparisons. The teacher-scored VA on the student's rollouts rises alongside accuracy in the reported VA-OPD training trajectories, consistent with stronger visual sensitivity alongside improved task accuracy.

Limitations. VA measures teacher-relative counterfactual sensitivity, which is a useful proxy for visual dependency but not a direct measurement of internal perceptual processes such as attention allocation or feature grounding. Its interpretation depends on the chosen teacher and image-degradation operator; degraded views can introduce distribution shift, motivating the operator and interpolation controls in Appendix B. Within visual understanding, the observed gains are larger on HallusionBench and AI2D than on MMStar; their size can vary with the required visual information and the student's existing capabilities.

Future work. VA-OPD could more directly complement RL-based visual perception methods: it prioritizes teacher-scored visual supervision during distillation, while RL optimizes task reward. Extending VA-OPD to region- or object-level visual dependency is also promising for future work.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their constructive feedback. This work was supported by the National Key Research and Development Program (2023ZD0120900) and the National Natural Science Foundation of China (62372454).

Competing interests. During the 36 months preceding submission, the authors had no additional financial relationships outside the submitted work that could be perceived to influence this work, beyond the affiliations listed in this paper.

References

[1] 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 *The twelfth international conference on learning representations*, 2024.

[2] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*, 2024.

[3] Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

[4] Jie Sun, Mao Zheng, Mingyang Song, Qiyong Zhong, Gengsheng Li, Zhepei Hong, Chang Wu, Pengfei Liu, Junfeng Fang, and Xiang Wang. EasyOPD: An easy-to-use on-policy distillation framework for large language models. *arXiv preprint arXiv:2607.11012*, 2026. URL https://arxiv.org/abs/2607.11012.

[5] Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-policy distillation with curriculum turn-level guidance for multi-turn agents. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

[6] Gengsheng Li, Mao Zheng, Mingyang Song, Jie Sun, Zeyuan Liu, Ruiqi Liu, Tianyu Yang, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Shiming Xiang, Jinqiao Wang, and Tat-Seng Chua. Data-free on-policy distillation: How far can we go without external data? *arXiv preprint arXiv:2609.14193*, 2026. URL https://arxiv.org/abs/2609.14193.

[7] Jie Sun, Mao Zheng, Mingyang Song, Zeyuan Liu, Gengsheng Li, Houcheng Jiang, Yilin Cheng, Bichuan Feng, Yuchen Cai, Junfeng Fang, and Xiang Wang. Distill what you trust: Reliability-aware multi-teacher on-policy distillation. *arXiv preprint arXiv:2609.23697*, 2026. URL https://arxiv.org/abs/2609.23697.

[8] Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026. URL https://arxiv.org/abs/2605.07725.

[9] 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, 2026. URL https://arxiv.org/abs/2605.18740.

[10] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

[11] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[12] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, 2021.

[13] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

[16] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025.

[17] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-Math: Does your large multimodal model achieve human-like mathematical reasoning?, 2024. URL `https://arxiv.org/abs/2407.01284`.

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

[19] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[20] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385, 2024.

[21] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016.

[22] 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*, pages 9556–9567, 2024.

[23] 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.

[24] 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, 2024.

[25] 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*, pages 11198–11201, 2024.

[26] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024.

[27] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[28] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in LVLMs. *arXiv preprint arXiv:2405.15683*, 2024.

[29] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273, 2024.

[30] Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025. URL https://arxiv.org/abs/2508.19652v1.

[31] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853, 2024.

[32] Jihao Gu, Yingyao Wang, Meng Cao, Pi Bu, Jun Song, Yancheng He, Shilong Li, and Bo Zheng. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation. *arXiv preprint arXiv:2412.14487*, 2024.

[33] Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*, 2025.

[34] Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, Zhaojie Liu, Qibin Hou, and Ming-Ming Cheng. Rethinking token-level policy optimization for multimodal chain-of-thought. *arXiv preprint arXiv:2603.22847*, 2026.

[35] Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. NoisyRollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025.

[36] Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking Kullback–Leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5737–5755, 2025.

[37] Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026. URL https://arxiv.org/abs/2603.07079v1.

[38] Yifan Ding, Xincheng Wei, Yoshua Y. Li, Ziheng Li, Yuquan Lu, Siyu Zhang, Dongsheng Ma, Rongxiang Weng, Xunliang Cai, and Yun Chen. SAF-OPD: Stable advantage fusion for on-policy distillation, 2026. URL https://arxiv.org/abs/2607.29209.

[39] Xincheng Wei, Yifan Ding, Yoshua Li, Yuquan Lu, Ziheng Li, Yi Lu, Dongsheng Ma, Rongxiang Weng, and Xunliang Cai. Better supervision is nearby: Neighborhood on-policy self-distillation, 2026. URL https://arxiv.org/abs/2609.39687.

[40] Run Yang, Runpeng Dai, Jie Sun, Jielei Zhang, Fan Zhou, Hongtu Zhu, Peiyi Li, and Longwen Gao. Influence-directed distillation: Solving the diversity bottleneck in sampled-token on-policy distillation. *arXiv preprint arXiv:2608.29846*, 2026. URL https://arxiv.org/abs/2608.29846.

[41] Jie Sun, Mao Zheng, Mingyang Song, Qiyong Zhong, Yilin Cheng, Bichuan Feng, Pengfei Liu, Junfeng Fang, and Xiang Wang. SimCT: Recovering lost supervision for cross-tokenizer on-policy distillation. In *Advances in Neural Information Processing Systems*, 2026.

[42] Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. *arXiv preprint arXiv:2410.11325*, 2024.

[43] Ximo Zhu, Ruiqi Liu, Rong Wang, Ping Wu, Xiang Zheng, Wenzhuo Xu, Xubin Yao, Zhiyuan Yan, Bo Li, Jun Gao, et al. ReOrder-OPD: Reliability-aware prompt ordering for on-policy distillation. *arXiv preprint arXiv:2608.10905*, 2026. URL https://arxiv.org/abs/2608.10905.

[44] Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/. Published on 27 October 2025.

[45] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

[46] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256.

[47] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

[48] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2024. URL https://arxiv.org/abs/2412.05271.

A Implementation Details

A.1 Training Configuration

We train with verl [46] using AdamW ($(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.1, learning rate 1×10^{-6}, cosine schedule with 5% linear warm-up) for 5 epochs. Each step uses 16 prompts with $K = 4$ rollouts per prompt; rollouts are sampled at temperature 0.7, top-p 0.95, with maximum length 4096. Evaluation uses temperature 1.0 with avg@8. VA-OPD hyperparameters: rollout-reweighting softmax temperature $\tau = 1.0$ and normalization stability $\epsilon = 10^{-6}$ (Equations (3) and (4)); token-grouping high-VA fraction $p_v = 0.2$ and high-VA-group weight $\lambda = 0.5$ (Equation (5)); pixelation ratio 0.10 for $\tilde{p}_{\mathrm{T}}$ (Appendix B). The additional forward pass on $\tilde{v}$ adds $\sim$7% wall-clock over Standard OPD on 8×A100 for the 8B→2B Geo3K run (114.3 vs. 106.4 s/step).

The primary Qwen3-VL-8B→2B Geometry3K training run uses a single 8×A100-80GB node in bfloat16 precision; rollouts are sampled with vLLM [47] colocated on the same node, and the teacher's pixelated forward pass shares the inference GPUs. We use the verl trainer interface but replace the policy-gradient loss with our reverse-KL objective. The student KL objective is evaluated on 16 prompts $\times$ K rollouts $\times$ T response-token positions per step, with student forward/backward computation and original-image teacher scoring; VA-OPD adds one extra teacher forward on $\tilde{v}$ for the VA estimate. We do not use gradient accumulation, FlashAttention is enabled by default in the Qwen3-VL implementation we build on, and rollouts that reach the 4096-token budget are truncated rather than discarded.

The time-to-accuracy comparison in Section 4.5 evaluates the MathVerse Vision Only subset, whose inputs contain minimal text questions and present the problem in the image.

A.2 Evaluation Protocol

All benchmarks follow their official evaluation protocols for consistency across methods. We evaluate with VLMEvalKit. The original matched setting caps each image at 2,560 visual tokens. Rule-based matching is used for WeMath, AI2D, MMMU, MMStar, and OCRBench; LLM-based scoring with GPT-4o as the judge (temperature 0) is used for MathVista, MathVerse, and HallusionBench. Following each benchmark's official scoring protocol, we report: WeMath as the average of the official Strict and Loose scores; HallusionBench as the average of aAcc, fAcc, and qAcc; and OCR-Bench as its official 1000-point Final Score normalized to a 0–100 scale (i.e., divided by 10). All other benchmarks (MathVista, MathVerse, AI2D, MMMU, MMStar) follow their default accuracy metrics. MathVerse and MathVista are evaluated on their testmini splits. Checkpoints are selected on 200 held-out Geometry3K problems disjoint from the training split.

For LLM-based scoring we pin the judge to GPT-4o version `gpt-4o-2024-08-06` and apply each benchmark's official judge prompt verbatim, with judge temperature 0 to reduce scoring variability. Each evaluation prompt is given to the student in its original released form, with no system-prompt rewriting and no extra chain-of-thought injection; the student's first committed answer is parsed and compared against the gold. For accuracy metrics, avg@8 averages binary per-sample correctness over 8 independently sampled rollouts at evaluation temperature 1.0, where per-sample correctness is 1 if the parsed answer matches the gold and 0 otherwise. Composite metrics retain their benchmark-specific aggregation. This averages correctness rather than selecting any successful answer or taking a majority vote. The same benchmark-specific parsing pipeline is reused across all baselines, keeping the answer-extraction policy consistent in Tables 1 and 2.

MathVerse evaluation scope. The original comparison in Tables 1 and 2 evaluates 788 inputs from `MathVerse_MINI_Vision_Only`, in which the problem is rendered in the image and the text question is minimal. We denote this subset by *MV-VO*. The complete `MathVerse_MINI` testmini consists of the same 788 problems in five versions—Text Dominant, Text Lite, Vision Intensive, Vision Dominant, and Vision Only—for 3940 evaluation inputs. The additional complete-testmini results in Tables 5 and 6 use all five versions, giving $3940 \times 8 = 31{,}520$ responses per method and configuration. These results supplement the original Vision Only comparison; scores from the two scopes are not combined in one benchmark average. The supplementary perturbation, interpolation, and non-Qwen MathVerse results below also use the complete testmini. MathVista uses its testmini split.

A.3 Sensitivity to p_v and λ

The two token-grouping hyperparameters of VA-OPD are the high-VA fraction p_v and the high-VA group weight λ (Equation (5)). Table 3 sweeps both over a 3×3 grid centered at the default $(p_v, \lambda) = (0.2, 0.5)$. MV-VO accuracy ranges from 30.6 to 31.9, a span of 1.3 points, and every configuration remains above the Standard OPD baseline of 29.1 from Table 1. With the default $\lambda = 0.5$, varying p_v from 0.1 to 0.3 changes accuracy by only 0.5 points. Ignoring integer group-size rounding, $p_v = \lambda = 0.3$ gives equal per-token coefficients in the two groups, so the objective reduces to rollout-level reweighting with uniform token KL weights; its 30.6 accuracy agrees with the rollout-reweighting-only result in Figure 4. These results support stability around the default while showing the benefit of giving high-VA positions a larger per-token loss coefficient.

Table 3: p_v **and λ sensitivity on MV-VO.**

	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$
$p_v = 0.1$	31.3	31.6	31.5
$p_v = 0.2$	31.7	**31.9**	31.4
$p_v = 0.3$	30.6	31.4	31.5

Table 4: **Pixelation-ratio sensitivity.** MathVerse is MV-VO.

Ratio	MV-VO	HalluB	OCRBench
0.05 (aggressive)	31.6	53.7	86.2
0.10 (default)	**31.9**	**54.5**	**86.4**
0.20 (mild)	31.3	54.1	85.7

A.4 CoT-SFT data construction

CoT-SFT filters teacher-generated trajectories from the shared training-question pool by final-answer agreement with the dataset annotations.

Geometry3K and ViRL39K do not provide complete natural-language reasoning trajectories. The corresponding teacher generates solutions from each training image and question; we retain trajectories whose parsed final answers match the ground truth, removing unparseable and overlength responses. CoT-SFT uses these fixed targets, whereas Standard OPD and VA-OPD supervise student-generated rollouts. The methods share the underlying training-question pool, but filtering can change the number of retained SFT examples.

A.5 Supplementary evaluation and generalization

The Geometry3K/ViRL39K comparison holds the teacher–student pair fixed while varying both the scale and example composition of the training corpus.

Table 5 evaluates the complete five-variant MathVerse testmini separately from the original MV-VO comparison. VA-OPD improves over Standard OPD by 0.8–2.4 points across the four reported configurations. Table 6 extends the fixed Qwen3-VL-32B teacher comparison to 4B and 8B students and tests InternVL2.5-8B→2B [48]. The larger-student and InternVL distillation runs use Geometry3K; student-size comparisons use matched training steps. The InternVL pair uses its official chat template and image preprocessing, the same rollout count, optimizer, and epochs as the primary setting, and the shared evaluation sampling/scoring protocol. These student/model-family tests cover complete MathVerse, HallusionBench, and OCRBench.

Table 5: **Complete MathVerse testmini evaluation.** Each configuration uses all five variants (Text Dominant, Text Lite, Vision Intensive, Vision Dominant, and Vision Only): 3,940 inputs and 31,520 responses per method at avg@8, temperature 1.0. These scores use the complete `MathVerse_MINI`, whereas Tables 1 and 2 report MV-VO.

Teacher→Student	Data	Base	OPD	VA-OPD
4B→2B	Geometry3K	51.7	52.7	**53.5**
8B→2B	Geometry3K	51.7	54.6	**56.3**
32B→2B	Geometry3K	51.7	57.2	**59.6**
8B→2B	ViRL39K	51.7	55.6	**57.9**

At `https://github.com/handsome-rich/VA-OPD`, training code, data preparation, and evaluation documentation are available.

Table 6: **Larger students and another VLM family.** All distillation runs use Geometry3K. Left: Qwen3-VL-32B teacher, with matched training steps across student sizes. Right: InternVL2.5-8B→2B with its official chat template and image preprocessing. MV-Full denotes the complete five-variant MathVerse testmini.

Student	Method	MV-Full	HalluB	OCR
2B	OPD	57.2	52.3	84.7
	VA-OPD	**59.6**	**55.1**	**86.9**
4B	OPD	59.0	58.7	88.8
	VA-OPD	**60.2**	**60.1**	**89.5**
8B	OPD	62.8	61.8	90.0
	VA-OPD	**63.7**	**62.5**	**90.4**

InternVL2.5	MV-Full	HalluB	OCRBench
Teacher (8B)	39.2	49.7	82.0
Base (2B)	30.1	42.4	80.1
OPD (2B)	33.2	44.5	81.1
VA-OPD (2B)	**35.1**	**45.7**	**81.9**

A.6 Sampling and optimization conventions

The student generates K rollouts $\{y^{(k)}\}_{k=1}^{K}$ using the decoding settings in Appendix A.1. These settings determine the sampled sequences, whereas p_{S} and p_{T} in Equations (1) and (5) denote the models' full-vocabulary next-token distributions used for KL supervision.

During each update, sampled sequences and prefixes are fixed observations, and teacher-derived weights and partitions are detached; gradients pass only through the student's full-distribution KL, without a score-function term for the rollout sampler.

A.7 Token-group weighting details

Rollout reweighting changes the allocation across sibling trajectories, while grouped KL changes the allocation across positions within each trajectory.

We retain both token groups rather than discarding low-VA positions. Ignoring integer rounding, the high-/low-group per-token coefficient ratio is $\lambda(1-p_v)/[(1-\lambda)p_v] = 4$ at the default $(p_v, \lambda) = (0.2, 0.5)$. At $\lambda = 0.5$, this default lies centrally within the empirically stable $p_v = 0.1$–0.3 interval (Appendix A.3): too small a high group can omit relevant positions, while a large group weakens their relative emphasis. For equal-sized groups, $p_v = \lambda = 0.5$ recovers uniform token averaging, leaving only rollout reweighting.

The released implementation makes its discrete grouping choices explicit: it selects $\lceil p_v T \rceil$ high-group positions and resolves equal VA values stably by token order by default. If the low group is empty, as for a one-token response, the high group receives the full rollout weight; empty responses are invalid. The accompanying software documentation specifies these grouping and edge-case conventions.

B Construction of the counterfactual teacher view

The teacher distribution $\tilde{p}_{\mathrm{T}}$ supplies the counterfactual logprobs that define visual advantage (Equation (2)); its construction directly affects the quality of the VA signal. We use pixelation, downsampling the image to 10% of each image dimension with bilinear interpolation and upsampling back with nearest-neighbor, which suppresses fine-grained visual detail while leaving a low-frequency approximation of the same scene. The broad scene structure and unchanged input dimensions provide a matched probe for the contribution of fine-grained information. The operator comparisons below evaluate how this controlled reduction in fine-grained detail affects VA concentration and downstream performance. The counterfactual view supplies VA log-probabilities rather than the reverse-KL target.

We compare pixelation against four widely considered alternatives: removing the vision token stream from the teacher input entirely; replacing the image with a uniformly black image of the same resolution; randomly permuting the image patches, which preserves low-level statistics but destroys layout; and randomly masking a fraction of image patches in pixel space (zeroing out the pixels of selected patches before the image is encoded). Each alternative is plugged into VA-OPD with no other changes and trained for the same number of steps; Table 7 reports the original-image downstream accuracies on MV-VO, HallusionBench, and OCRBench under the same setting as

Table 7: **Pixelation vs. alternative constructions of $\tilde{p}_{\mathrm{T}}$.**

Construction of $\tilde{p}_{\mathrm{T}}$	MV-VO	HalluB	OCRBench
Standard OPD (no $\tilde{p}_{\mathrm{T}}$)	29.1	52.0	84.7
Remove vision tokens	28.8	51.7	84.5
Blank (black) image	29.0	51.5	84.6
Patch-shuffle	28.9	52.0	84.3
Random patch mask	29.2	51.4	84.8
Pixelation (ours)	**31.9**	**54.5**	**86.4**

Figure 4, with Standard OPD as the reference. Only pixelation produces consistent gains over the OPD baseline across all three benchmarks; the other four constructions remain near OPD or fall slightly below it, showing that degradation choice affects the gains among these tested alternatives. Within the pixelation family, varying the downsample ratio (Table 4) leaves downstream accuracy stable across the 0.05–0.20 range.

On fixed held-out Standard OPD rollouts across complete MathVerse, HallusionBench, and OCR-Bench, pixelation places 91.8–96.6% of VA mass in the top 20% of positions, compared with 50.2–76.9% for patch shuffle, random patch masking, and blank images. Bilinear/nearest-neighbor, bicubic/nearest-neighbor, and area/bilinear interpolation obtain complete-MathVerse scores of 56.3, 56.4, and 56.1, respectively. Together, the concentration and interpolation results support pixelation as a counterfactual construction that concentrates VA and maintains downstream accuracy across the tested interpolation choices.

C Independent validation of visual reliance

Independent checks complement the qualitative VA heatmap in Figure 6. In 100 OCRBench responses, two authors manually verified image-dependent spans proposed by `gemini-2.5-flash` and aligned them to model tokens. These tokens constitute 17.2% of all output positions but 68.7% of the top-20% VA set. On original-image MathVerse Vision Only responses, high-VA positions also receive more image attention than the remainder: the mean gaps are 0.087, 0.035, and 0.027 for VA-OPD, Standard OPD, and Base, respectively. This enrichment and attention association support VA as a proxy for image-influenced generation positions; the following paragraphs give their protocols and additional behavioral controls.

For original-image MV-VO responses, we sum each output token's attention over visual positions, average over layers and heads, compare each response's top-20% VA group with its complement, and average over responses. Table 8 reports the resulting attention. VA determines the groups, and attention is measured separately on each model's own generated responses. The comparison characterizes the association between teacher-derived sensitivity and the model's image-attention pattern.

For 100 OCRBench responses, `gemini-2.5-flash` proposes spans corresponding to image-derived target text or numbers absent from the question. Two authors check every annotation and align the verified spans to model tokens. These annotations are defined independently of VA. Image-dependent tokens occupy 17.2% of all positions and 68.7% of the high-VA subset, an approximately fourfold enrichment.

Table 8: **Independent token-level validation.** Left: mean attention to image tokens on each model's original-image MV-VO responses, comparing top-20% VA tokens with the remainder. Right: manually verified OCRBench token categories in 100 responses; percentages give their shares of all output tokens and the high-VA subset.

Model	High-VA	Remaining
Base	0.181	0.154
OPD	0.204	0.169
VA-OPD	**0.263**	0.176

OCR token category	All	High-VA
Image-dependent	17.2%	**68.7%**
Other	82.8%	31.3%

As a complementary behavioral check, Table 9 evaluates Base, Standard OPD, and pixelation-trained VA-OPD on shuffled or randomly masked images. These operators are used only at evaluation for the compared models. VA-OPD's margin over Standard OPD narrows from 1.7–2.5 points on original images to 0.4–0.9 under perturbations. Its larger gains on intact images are consistent with the method's emphasis on fine-grained visual information.

Table 9: **Held-out evaluation perturbations.** Each cell reports Base / Standard OPD / VA-OPD. MathVerse uses all five testmini versions.

Image input	MathVerse	HalluB	OCRBench
Original	51.7 / 54.6 / 56.3	52.6 / 52.0 / 54.5	85.9 / 84.7 / 86.4
Patch shuffle	38.1 / 39.4 / 40.3	36.6 / 36.8 / 37.2	63.5 / 64.0 / 64.7
Random patch mask	44.8 / 46.5 / 47.1	45.0 / 44.7 / 45.5	76.3 / 76.2 / 77.1

High-VA function words. In Figure 6, *given* is strongly shaded before the image-derived coordinates in "The point P is given as $(-21/29, 20/29)$". The words *for* and *an* are shaded in "In standard position, for an angle a, the tangent of the angle is defined as ...". VA measures teacher-probability changes in context: here the tangent definition supplies a general mathematical rule, and *a* names the angle variable. The heatmap displays continuous VA scores, while top-20% membership is determined by ranking all positions within the response.

Validation coverage. The eight-benchmark study uses Qwen3-VL with a 2B student, while the additional student/model-family tests cover three benchmarks and the OCR annotation covers 100 responses. Together, token annotations, image-attention measurements, and evaluation perturbations provide complementary checks of image-influenced generation at the token and response levels.

D Relation to perception-aware reinforcement learning

VPPO [33] also identifies sparse token dependence and variation across trajectories. It compares the policy's predictive distributions under original and patch-masked images, retains reward-based policy gradients on the top-$k\%$ dependency positions, and min–max normalizes trajectory dependency to shape advantages. VA-OPD instead scores realized student tokens with an external teacher under original and pixelated images. Its softmax weights sibling rollouts, while separately normalized high- *and* low-VA groups retain full-distribution reverse-KL supervision. This distinction concerns the supervision source and optimization objective, beyond sharing a two-level treatment of visual dependence.

E Additional Qualitative Examples

Figures 7 and 8 present nine additional qualitative cases from MathVista, MathVerse, AI2D, HallusionBench, MMMU, and MMStar. Each row uses the same input and prompting setup for three students: Base, Standard OPD, and VA-OPD, enabling a response-level comparison. For each student, the figure reports the rollout-averaged VA, the maximum per-token VA, and a summarized output marked as Wrong, OK, or No final. Wrong and OK indicate whether the parsed final answer agrees with the benchmark label; No final indicates that no parseable final answer is given. The response summaries accompany these final-answer labels with the reasoning and visual descriptions expressed by each student. For VA-OPD, the listed high-VA tokens identify positions with stronger teacher-scored sensitivity to fine-grained visual detail.

Base and Standard OPD generally have lower rollout-level VA means than VA-OPD on the same examples. Their summarized outputs include missed objects, incorrect label mappings, and failures to produce a stable final answer. In the AI2D brain case, both describe the region labeled D as medulla, whereas the diagram points to cerebellum. The MMMU protein row likewise shows a direct answer-level difference: both select region B, while the reference answer is region A, the mostly α-helical part separated by the dashed boundary. In the Hallusion-Simpsons row, Standard OPD and VA-OPD both answer No, while their summaries differ in character naming and action description. This contrast highlights response-level differences even when the parsed final answers coincide, complementing cases where answer correctness changes across students.

In the VA-OPD columns, the listed high-VA tokens include numbers, entities, and geometric relations appearing in the tasks. Several displayed examples have higher rollout-level VA means than Standard OPD, including 0.11 vs. 0.05 on the AI2D brain case, 0.21 vs. 0.08 on the MathVerse function-graph case, and 0.32 vs. 0.23 on the MMStar soccer-activity case. The brain case lists cerebellum / label D / brain / diagram, while the Chinese parallel-lines case lists the two lines and their transversal. In the latter, VA-OPD gives 118° using the supplementary-angle relation, while Base and Standard OPD produce no final answer. Its displayed answers agree with the benchmark labels on all nine cases, including the protein and function-graph rows. The soccer row likewise pairs benchmark-label agreement with higher mean VA, extending the response-level comparison to an activity question alongside the object- and relation-focused examples. On the band-counting row, both trained students report four people, including the drummer; the VA-OPD high-VA list includes drummer and seated. Computed under each response's own prefix, these VA statistics characterize the teacher's predictive sensitivity at the listed positions. The systematic analyses in Appendix C complement this response-level comparison with independently measured image attention and verified image-dependent token categories.

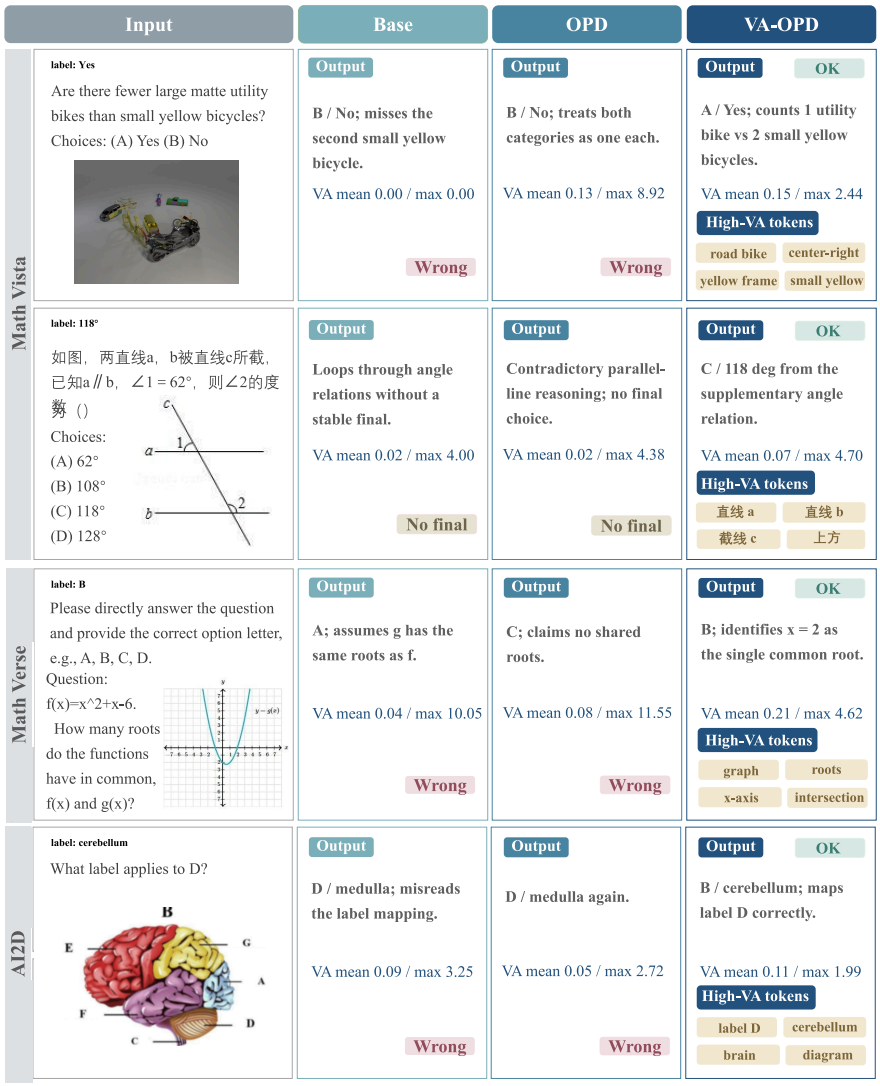


Figure 7: **Additional qualitative case 1.** Base, Standard OPD, and VA-OPD on the same inputs from MathVista, MathVerse, and AI2D. Each cell summarizes the response, its mean and maximum token-level VA, and parsed-answer agreement with the benchmark label (Wrong, OK, or No final); the VA-OPD column also identifies high-VA positions.



Figure 8: **Additional qualitative case 2.** HallusionBench, MMMU, and MMStar examples. Further comparisons of Base, Standard OPD, and VA-OPD on the same image–question pairs. The reported response summaries, VA statistics, correctness labels, and high-VA positions follow the conventions of Figure 7.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The original eight-benchmark comparisons are reported separately from complete-MathVerse reevaluation and additional student/model-family tests (Tables 1, 2, 5 and 6). Attention comparisons, OCR annotation, and held-out perturbation tests evaluate the teacher-relative image sensitivity measured by VA (Section 5 and Appendix C).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sections 2.1 and 7 and Appendix B discuss the dependence of VA on the teacher and degraded-image probe. Appendix C describes the scope of perturbation tests, attention comparisons on separately generated responses, and OCRBench-specific annotation validation; complete-MathVerse and original MV-VO results are reported separately.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical; the equations in Sections 2 and 3 are method definitions (visual advantage, rollout-level reweighting, token-level grouped KL, the overall objective) rather than theorems requiring proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The primary training recipe and evaluation protocol are provided in Appendices A.1 and A.2, and the public code implements Standard OPD and VA-OPD. CoT-SFT target-generation settings and some comparison-baseline configurations are not fully specified.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way

to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The public package at `https://github.com/handsome-rich/VA-OPD` includes Standard OPD/VA-OPD training, data preparation, and score aggregation using public source models and benchmarks. Complete reproduction scripts for the other baselines, the InternVL experiment, and all additional diagnostics are not included; the additional experiments are reported in Appendix A.5.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: Section 4.1 and Appendices A.1 and A.2 specify the main training recipe, score definitions, original image budget, and MathVerse scopes. Some experiment-specific settings, including the full CoT-SFT generation/filtering and diagnostic configuration, are not completely reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report avg@8, averaging eight independently sampled responses per question at evaluation time. Confidence intervals, significance tests, and variance across independent training seeds are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix A.1 reports the primary eight-A100-80GB setup and 114.3 versus 106.4 seconds per step; Section 4.5 reports time to a fixed MV-VO accuracy. We do not provide a full per-experiment resource/time breakdown or total compute for the student-scaling, InternVL, and diagnostic runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly released models and benchmarks, involves no human subjects, and we have reviewed and conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper studies a distillation methodology for vision-language models with no application or deployment claim; we do not identify direct positive or negative societal impacts beyond those already attached to VLM research in general.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The new release is a training-code package. The reused Qwen3-VL and InternVL2.5 checkpoints and benchmark datasets remain subject to their original distribution terms.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The principal model and benchmark sources are cited. The public code package provides an Apache-2.0 license and retained third-party notices, but this paper does not enumerate the version-specific licenses and terms for every source checkpoint and dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The public code package at `https://github.com/handsome-rich/VA-OPD` includes installation and training instructions, data-preparation documentation, an evaluation workflow, an Apache-2.0 license, and a description of experiment coverage.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Two authors conduct the OCRBench annotation checks described in Appendix C. The study includes no crowdsourcing or human-participant experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study does not conduct experiments on human participants. The manual check concerns generated answers and image-dependent token spans in an existing benchmark.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The method uses a pretrained VLM teacher for token probabilities, and the CoT-SFT baseline uses teacher-generated trajectories (Appendix A.4); GPT-4o judges MathVista, MathVerse, and HallusionBench, while rule-based matching scores WeMath, AI2D, MMMU, MMStar, and OCRBench (Appendix A.2). Gemini-2.5-Flash proposes OCR token spans that two authors manually verify (Appendix C).

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.