OPD-AHA: FROM LINGUISTIC MOMENTUM TO VISUAL REFLECTION IN MULTIMODAL ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Privileged on-policy distillation improves multimodal reasoning by allowing a teacher to evaluate student trajectories using rich, training-only visual evidence. Both models score these trajectories while conditioning on the same student-generated prefix. When a student misinterprets an image early in a response, this accumulating erroneous rationale eventually pulls the teacher away from its visual evidence. The teacher and student converge on the same hallucination, causing standard cross-model supervision to collapse precisely where correction is most needed. We find that the teacher's visual corrective preference is not lost under this misleading agreement. Comparing the predictions of the identical teacher given the real image and a visual null reveals that the privileged evidence still pushes the model toward the correct interpretation. We introduce OPD-Aha, which reconstructs the distillation target directly from this isolated visual preference rather than relying on the fragile teacher-student discrepancy. This reconstructed target aggressively suppresses continuations that contradict the image. Trained with this objective, students learn to naturally interrupt their own flawed reasoning with reflection tokens such as *wait* and *actually*. After reflection, subsequent generation relies less on the accumulated erroneous text and more on the visual evidence. Correcting these trajectories mid-generation fundamentally alters the reasoning process, yielding broad and consistent improvements across diverse fine-grained perception and complex multimodal reasoning benchmarks. Our code and models are available at `https://anonymous.4open.science/r/OPD-Aha-ananomy-CC87`.

1 INTRODUCTION

On-policy distillation provides dense, token-level supervision on the exact trajectories explored by a student model (Agarwal et al., 2024; Gu et al., 2024; Zhao et al., 2026; Jin et al., 2026; Li et al., 2026a). In multimodal reasoning (Liu et al., 2023; Li et al., 2023a; Bai et al., 2023), privileged on-policy distillation strengthens this supervision by giving the teacher access to richer, training-only visual evidence (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016), such as localized high-resolution views, while the student continues to operate on its original visual input (Yuan et al., 2026; Tian et al., 2026; Wei et al., 2026). This privileged evidence is used to evaluate the states visited by the student during its own rollout. Accordingly, at every decoding step, the teacher combines its richer visual input with the same student-generated linguistic prefix that defines the current student state. As the rollout progresses, privileged visual supervision is therefore delivered under an increasingly long context written by the student itself.

This coupling becomes problematic when the student makes an early perceptual error (Li et al., 2026d). Once this error enters the prefix (Jiang et al., 2026; Xu et al., 2026), subsequent generation elaborates on an interpretation that conflicts with the image (Li et al., 2023b; Favero et al., 2024; He et al., 2025; Guo et al., 2025b; Chen et al., 2026b). The teacher must then evaluate its privileged visual evidence in the presence of an increasingly strong linguistic context supporting the student's mistaken interpretation. We show that as this erroneous text grows, its linguistic momentum progressively dominates the teacher's predictions and marginalizes the visual evidence. The teacher eventually abandons the visually grounded correction and favors the student's hallucinated contin-

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

uation. Standard privileged distillation therefore loses its corrective signal precisely at the states where the student most needs intervention.

Despite this apparent supervision collapse, we find that a robust visual corrective preference survives in the teacher's predictions. We introduce OPD-Aha to reconstruct the distillation target directly from this surviving signal. The method exposes the hidden visual preference by evaluating the identical teacher under the same shared prefix, varying only the visual input between the privileged image and a visual null (Leng et al., 2024; Favero et al., 2024). This intra-teacher contrast strips away the linguistic momentum and isolates the pure effect of the visual evidence, revealing that the privileged evidence continues to strongly suppress the hallucinated continuation even when the teacher's overall token distribution aligns with the student's erroneous reasoning. OPD-Aha translates this real-null prediction difference into a regularized target distribution that selectively suppresses image-inconsistent continuations while preserving a valid language distribution from the privileged teacher. Distilling this reconstructed target along the unchanged student rollout equips the student with a mechanism to interrupt and re-anchor its generation on visual evidence.

Students trained with OPD-Aha learn to naturally interrupt their own flawed reasoning, producing reflection tokens such as *wait* and *actually* (Guo et al., 2025a; Zhou et al., 2025) when their current explanation conflicts with the image. We find that this behavior emerges not because the reconstructed target directly increases the absolute probability of reflection tokens, but because it suppresses the erroneous continuation more strongly. This asymmetric suppression grants self-interruption a crucial relative advantage precisely where correction is needed. We further examine how this self-interruption alters the generation dynamics. After reflection, the student relies less on its earlier erroneous explanation and draws more strongly on the visual evidence when generating subsequent tokens. This restored visual reliance yields consistent accuracy improvements across six fine-grained perception and complex multimodal reasoning benchmarks.

Our contributions are as follows:

1. We identify a failure mode of privileged on-policy distillation: erroneous student prefixes can overwhelm privileged visual evidence, collapsing teacher–student supervision precisely when correction is most needed.
2. We introduce OPD-Aha, which reconstructs supervision from an intra-teacher real–null contrast under the same shared student prefix. This contrast isolates a corrective visual preference that survives the collapse and enables self-interruption with renewed visual reliance.
3. Training with OPD-Aha yields consistent accuracy improvements across fine-grained perception benchmarks and positive transfer to unseen multimodal reasoning tasks. More broadly, our results show that robust multimodal distillation requires preserving visual correction against the linguistic momentum of the student trajectory.

2 Student Prefixes Mask Privileged Supervision

2.1 Privileged Multimodal On-Policy Distillation

Privileged multimodal on-policy distillation exploits an asymmetry in visual evidence between the teacher and student. Given a standard visual observation I and a textual query x, the trainable student p_θ generates a reasoning trajectory $y \sim p_\theta(\cdot \mid I, x)$. A frozen teacher p_ϕ evaluates the student-generated states while receiving additional training-only visual evidence I^+, such as a localized high-resolution view of the task-relevant region, that is unavailable to the student.

Despite this visual asymmetry, the teacher and student are conditioned on the same student-generated linguistic prefix $h_t = (x, y_{<t})$ at every decoding step. Standard privileged multimodal OPD (Yuan et al., 2026) minimizes the token-level divergence between their predictions over these shared contexts:

$$\mathcal{L}_{\mathrm{MOPD}}(\theta) = \mathbb{E}_{y \sim p_\theta(\cdot \mid I, x)} \left[\frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{dist}}\big(p_\phi(\cdot \mid I^+, h_t), p_\theta(\cdot \mid I, h_t)\big) \right]. \tag{1}$$

Privileged visual evidence gives the teacher access to stronger visual grounding than the student. Standard OPD transfers this advantage through the teacher–student prediction discrepancy under

their shared linguistic prefix. We next examine how this linguistic momentum affects privileged supervision when the student's trajectory already contains an erroneous visual interpretation.

2.2 Erroneous Student Prefixes Mask Privileged Supervision

We find that an erroneous student prefix can progressively override the teacher's privileged visual evidence. Although the teacher receives stronger visual evidence, its autoregressive predictions remain conditioned on the same student-generated prefix, which may already encode an interpretation that contradicts the image. To quantify this effect, we retain progressively longer portions of failed Vision-OPD trajectories (Yuan et al., 2026) and measure the teacher's preference between the correct answer and the student's realized incorrect answer under each retained prefix. Their difference in length-normalized log-likelihood defines a decision margin that indicates whether the teacher still favors a visually grounded correction or the erroneous prefix has become dominant.

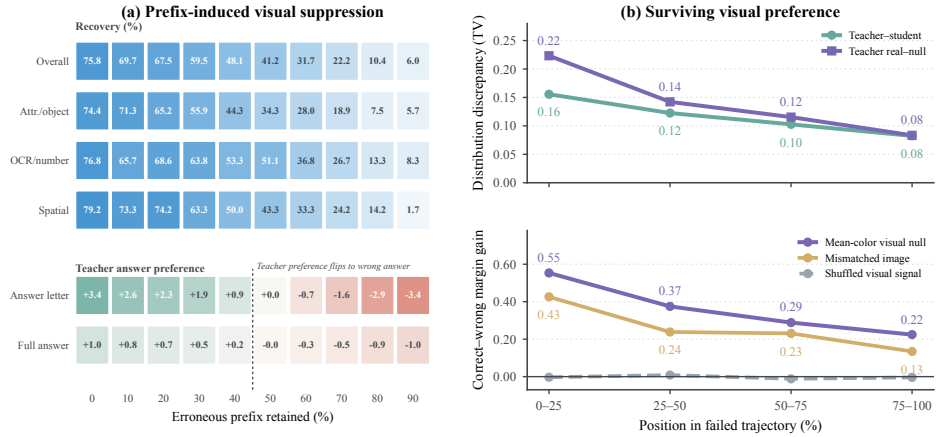


Figure 1: Erroneous prefixes mask privileged visual supervision. **(a)** As failed reasoning accumulates, the teacher's preference flips from the correct to the wrong answer. **(b)** The teacher–student discrepancy collapses, yet an intra-teacher visual contrast retains a strong preference for the correct answer. Token shuffling destroys this signal, confirming a token-specific visual preference.

The privileged teacher reliably recovers the correct answer from short prefixes, but this ability drops sharply as erroneous reasoning accumulates (Figure 1a). Near the midpoint of the response, its decision margin changes from positive to negative. The teacher no longer favors a visually grounded correction and instead prefers the student's realized wrong answer. As both models follow the same erroneous branch, their predictions converge and the cross-model discrepancy that drives standard privileged OPD disappears precisely where correction is most needed.

This convergence creates the impression that visual evidence has ceased to matter. We test this possibility by comparing two changes under the same student prefix: the difference between the teacher and student predictions, and the difference between the same teacher evaluated with privileged evidence and a visual null. The cross-model discrepancy collapses, while the real–null change remains substantial and continues to favor the correct answer (Figure 1b). Replacing the visual null with a mismatched natural image preserves this direction, whereas shuffling the visual change across tokens destroys it. The surviving response is therefore a token-specific visual preference toward correction rather than undirected sensitivity to the input.

3 Reconstructing Visual Supervision

We introduce OPD-Aha to reconstruct the visually grounded supervision that standard privileged OPD loses under an erroneous student prefix. As summarized in Figure 2, OPD-Aha abandons the unreliable cross-model comparison and instead extracts the surviving visual preference by contrasting the privileged teacher's predictions under real evidence and a visual null (Section 3.1). Then, it translates this isolated preference into a KL-regularized distillation target that selectively suppresses

image-inconsistent continuations while preserving a valid language distribution (Section 3.2), such that supervision over the reconstructed target transfers the visual correction along the intact rollout.

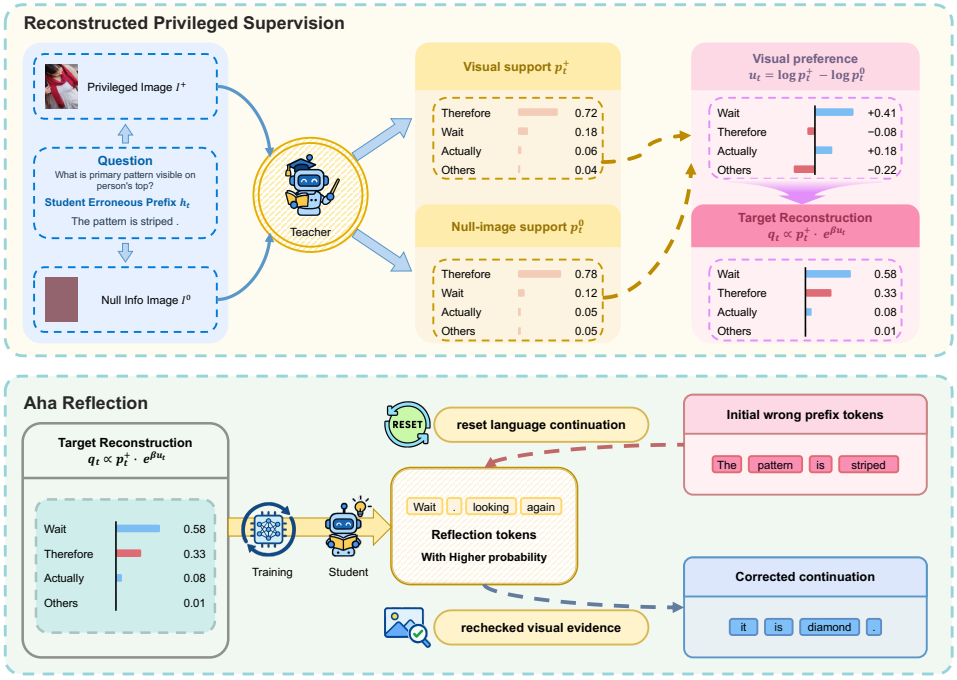


Figure 2: Overview of the OPD-Aha framework.

3.1 Isolating Visual Preference

Teacher–student agreement under an erroneous prefix does not imply that visual evidence has stopped affecting the teacher. As shown in Section 2, the cross-model discrepancy can collapse while a visual corrective preference remains. To expose this surviving signal without confounding it with differences between teacher and student, we compare the same teacher under the same student prefix while varying only its visual input (Yang et al., 2024; Zhao et al., 2025). Specifically, we pair the privileged image I^+ with a visual null

$$I^0 = \mathcal{N}(I^+), \tag{2}$$

where $\mathcal{N}$ removes informative visual content while preserving the input dimensions. We instantiate $\mathcal{N}$ by replacing the image with its mean RGB color. Under the same student prefix h_t, we denote the teacher distributions under the privileged image and visual null, and the student distribution under the original image, respectively, as

$$p_t^+(v) = p_\phi(v \mid I^+, h_t), \quad p_t^0(v) = p_\phi(v \mid I^0, h_t), \quad p_t^S(v) = p_\theta(v \mid I, h_t). \tag{3}$$

Since the teacher and linguistic prefix are held fixed between p_t^+ and p_t^0, their difference isolates how privileged visual evidence changes the teacher's token preferences. We define this difference as

$$u_t(v) = \log p_t^+(v) - \log p_t^0(v). \tag{4}$$

where a positive $u_t(v)$ means that privileged visual evidence increases the teacher's relative preference for token v, whereas a negative value means that the evidence suppresses it. Importantly, u_t can remain substantial even when p_t^+ and p_t^S are nearly indistinguishable, thereby revealing visual information that is hidden by teacher–student agreement.

Because p_t^+ and p_t^0 differ only in their visual input, u_t isolates the token-level effect of privileged visual evidence. The resulting visual preference does not itself encode correctness: a token favored by the image is not necessarily a correct continuation (Yin et al., 2025). Its corrective value instead lies in how it changes the relative preference between competing continuations. In particular, when visual evidence suppresses tokens that sustain an erroneous interpretation more strongly than alternatives, u_t points away from the hallucinated continuation.

3.2 Target Reconstruction from Visual Preference

The visual preference specifies how the teacher's token probabilities should shift in response to visual evidence, but it is not itself a probability distribution from which to supervise the student. We therefore use the privileged teacher distribution p_t^+ as an anchor and reconstruct a target that moves along the visual-preference direction while remaining close to the teacher's original distribution. Formally, we seek a target $q \in \Delta(\mathcal{V})$ that maximizes alignment with the visual preference while regularizing its departure from p_t^+:

$$q_t = \arg\max_{q \in \Delta(\mathcal{V})} \left\{ \beta \, \mathbb{E}_{v \sim q}[u_t(v)] - D_{\mathrm{KL}}\big(q \,\|\, p_t^+\big) \right\}. \tag{5}$$

The coefficient $\beta \geq 0$ controls the strength of the visual correction. Setting $\beta = 0$ leaves the base teacher distribution unchanged and recovers standard privileged OPD. Solving this objective applies an exponential tilt to the privileged target:

$$q_t(v) = \frac{p_t^+(v) \exp(\beta u_t(v))}{\sum_{w \in \mathcal{V}} p_t^+(w) \exp(\beta u_t(w))}. \tag{6}$$

Substituting the definition of u_t expands this solution into its component distributions:

$$q_t(v) = \mathrm{softmax}\left((1+\beta) \log p_t^+ - \beta \log p_t^0 \right)_v. \tag{7}$$

This expanded form reveals the mechanics of target reconstruction. The base probability p_t^+ preserves the teacher's complete belief under real evidence, ensuring that the target remains a valid language distribution. The likelihood ratio $(p_t^+ / p_t^0)^\beta$ then selectively amplifies or suppresses tokens based on how the visual evidence changes the teacher's preference. The target moves away from standard privileged OPD only along the directions favored by the visual input.

The effect of this selective amplification becomes clear when evaluating the relative odds of two competing tokens v and w:

$$\log \frac{q_t(v)}{q_t(w)} = \log \frac{p_t^+(v)}{p_t^+(w)} + \beta\big(u_t(v) - u_t(w)\big). \tag{8}$$

This competition governs whether the student continues its current explanation or interrupts itself. If w is an image-inconsistent continuation heavily favored by the inherited prefix, the base teacher preference $\log(p_t^+(v)/p_t^+(w))$ will strongly support w. Reconstruction can overturn this preference and promote a reflection token v if the visual evidence suppresses the continuation strongly enough to make the second term dominant. The parameter β determines how much visual separation is required to cross this threshold (Appendix E).

The reconstructed target q_t replaces the standard privileged teacher in the distillation divergence:

$$\ell_t = D_{\mathrm{dist}}\big(q_t, p_t^S\big). \tag{9}$$

The training objective averages this loss over all valid response positions M_t:

$$\mathcal{L}_{OPD-Aha} = \frac{\sum_t M_t \, \ell_t}{\sum_t M_t}, \tag{10}$$

The student optimizes this objective along its own unchanged rollout. The visual preference is transferred entirely through the adjusted target probabilities, allowing the student to learn visually grounded corrections without requiring auxiliary visual inputs during inference.

4 Emergent Reflection from Reconstructed Supervision

4.1 Reconstruction Sustains Corrective Supervision

We test whether target reconstruction preserves support for the correct answer as erroneous reasoning accumulates. Starting from failed student responses, we retain progressively longer prefixes

and compare the correct-answer probability under the standard privileged target p_t^+ and the reconstructed target q_t at the same prefix state. We also vary reconstruction strength β during training and track response length and final benchmark accuracy to examine how changes in correct-answer probability relate to the student's generation and performance.

Under the same short prefix, we find that both targets assign similar probabilities to the correct answer. As the erroneous prefix accumulates, standard privileged supervision sharply abandons the correct answer. The reconstructed target instead isolates the surviving visual preference, maintaining over an order of magnitude higher correct-answer probability at late-prefix states (Figure 3a,b). Support for the correct answer therefore remains stronger under the reconstructed target even after the student's trajectory has deviated.

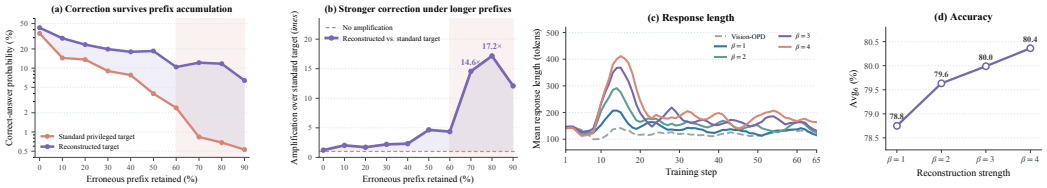


Figure 3: Late-prefix target recovery and training trends. **(a, b)** As hallucinated text accumulates, standard privileged distillation abandons the correct answer, whereas target reconstruction isolates the surviving visual preference to selectively amplify the correct token. **(c, d)** Stronger reconstruction induces a larger transient increase in response length before stabilizing, ultimately yielding consistent improvements in final accuracy.

This sustained visual correction alters the student's generation dynamics during training. Students trained with stronger reconstruction exhibit a larger transient increase in response length before their trajectories stabilize (Figure 3c). Across the same range of reconstruction strengths, overall accuracy averaged across the evaluated visual benchmarks improves consistently (Figure 3d). Preserving support for the correct answer along erroneous prefixes is thus accompanied by changes in generation and improved final performance.

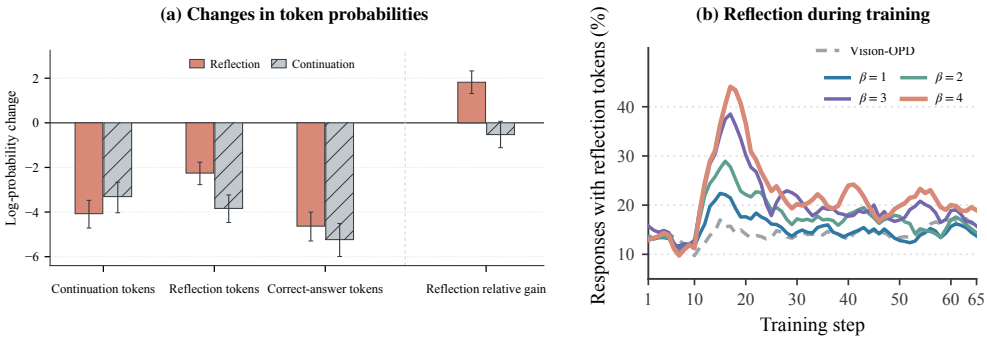


Figure 4: Token probabilities and reflection during training. **(a)** Target reconstruction selectively suppresses erroneous continuations before reflection, granting reflection tokens a relative advantage absent at non-reflection positions. **(b)** Stronger reconstruction increases the frequency of reflection tokens during training before stabilizing.

4.2 Suppressing Erroneous Continuations Elicits Reflection

We examine whether reconstruction promotes reflection by raising its probability or suppressing the erroneous continuation more strongly. We compare how reconstruction changes token probabilities at positions immediately before reflection and at matched positions in responses without reflection (Figure 4a). For each token category, we measure the log-probability change from p_t^+ to q_t, using its total probability. We evaluate the relative log-probability gain of reflection tokens over continuation

tokens to quantify the competitive advantage of self-interruption. To examine how these local probability adjustments translate into generation behavior, we track the overall frequency of reflection tokens throughout training across reconstruction strengths.

At these pre-reflection states, we find that reflection gains a relative advantage because reconstruction suppresses the erroneous continuation more strongly, even as the total probability of reflection tokens decreases. The correct-answer probability also decreases at these positions (Figure 4a). The visual preference thus discourages continuing the image-inconsistent explanation without directly rewarding reflection. Matched positions without reflection do not show the same increase in reflection tokens' relative probability.

This relative advantage is accompanied by more frequent reflection during training. As reconstruction strength increases, reflection tokens become more frequent during the transient phase and remain elevated for the stronger settings after stabilization, while Vision-OPD shows no comparable transition (Figure 4b). The student thus learns to interrupt its explanation when visual evidence reduces support for continuing it.

4.3 Reflection Restores Visual Reliance

We examine whether reflection is followed by a shift from textual to visual reliance by aligning generated responses at their first reflection token and comparing subsequent tokens with matched positions in responses without reflection. At each position, we measure predictive support from the accumulated student prefix and support gained from visual evidence, tracking how their balance evolves after self-interruption.

We observe that this balance shifts toward visual evidence after reflection compared with matched positions in responses without reflection. Support from the accumulated student prefix decreases, followed by an increase in visual support after a short delay (Figure 5). Reflection therefore marks a transition toward visual evidence, with a temporal gap between self-interruption and the recovery of visual support.

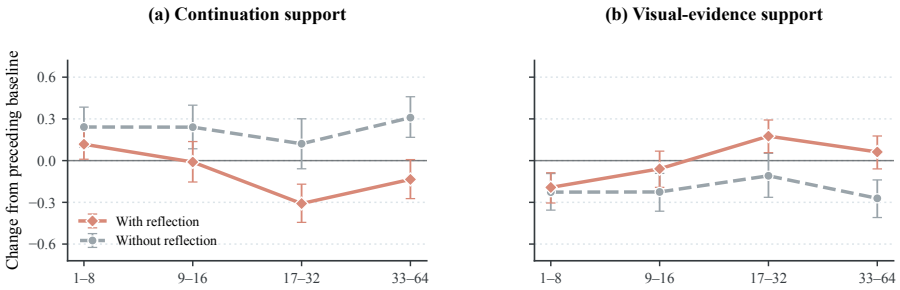


Figure 5: Visual reliance after reflection. **(a)** Support from the accumulated student prefix decreases after reflection. **(b)** Support gained from visual evidence rises after a short delay.

5 Experiments

5.1 Performance on Fine-Grained Visual Reasoning

Reconstructing the distillation target from the teacher's visual preference consistently translates into stronger downstream visual reasoning. OPD-Aha outperforms Vision-OPD across all six fine-grained, high-resolution, and real-world benchmarks at both evaluated model scales (Table 1). Rather than relying on fragile cross-model discrepancies, the reconstructed supervision provides a broad advantage, yielding a 4.0% absolute increase in accuracy at 4B and a 3.2% increase at 9B.

5.2 Comparison of Reconstruction Signals

We compare targets reconstructed from the standard teacher–student discrepancy ($\log p_t^+ - \log p_t^S$) and the intra-teacher real–null difference ($\log p_t^+ - \log p_t^0$) to verify that the downstream gains orig-

Table 1: OPD-Aha consistently outperforms standard privileged distillation and zero-shot baselines across diverse visual reasoning benchmarks. All OPD-Aha results use $\beta = 4$.

Models	Size	Fine-Grained		High-Resolution		Real-World		Avg_6
		$V^\star$	Zoom	HR-4K	HR-8K	MME-EN	MME-CN	
General-Purpose Models								
Gemini 3.1 Pro (Google DeepMind, 2026)	–	88.2	61.6	88.8	85.8	75.1	72.5	78.7
GPT-5.4 (OpenAI, 2026)	–	81.4	55.1	85.3	77.9	74.2	70.8	74.1
Kimi-K2.6 (Moonshot AI, 2026)	1T	86.3	53.8	82.7	78.7	69.7	66.4	72.9
"Thinking-with-Images" Agentic Models								
DeepEyes (Zheng et al., 2026)	7B	82.6	46.2	75.5	70.0	64.0	62.4	66.8
Thyme-RL (Zhang et al., 2025a)	7B	78.5	46.2	78.9	70.9	64.2	61.4	66.7
DeepEyesV2 (Hong et al., 2026)	7B	78.0	46.1	79.8	72.6	64.3	61.9	67.1
SenseNova-MARS (Chng et al., 2026)	8B	88.4	49.0	85.0	77.3	67.3	65.7	72.1
Qwen3.5 (Qwen Team, 2026)	4B	82.7	49.1	86.5	81.5	59.8	61.1	70.1
GRPO (Shao et al., 2024)	4B	85.2	57.3	78.7	75.4	70.9	68.5	72.7
V-Zero (Sun et al., 2026)	4B	78.0	56.5	84.8	80.9	73.2	71.1	74.1
Vision-OPD (Yuan et al., 2026)	4B	89.0	59.5	83.4	80.5	74.6	71.6	76.4
OPD-Aha (Ours)	4B	**93.7**	**62.5**	**88.6**	**84.8**	**77.7**	**74.9**	**80.4**
Qwen3.5 (Qwen Team, 2026)	9B	85.3	52.2	86.0	81.6	71.3	67.5	74.0
GRPO (Shao et al., 2024)	9B	87.2	57.5	86.0	83.0	73.3	69.2	76.0
V-Zero (Sun et al., 2026)	9B	89.2	59.4	85.8	83.3	75.7	70.6	77.3
Vision-OPD (Yuan et al., 2026)	9B	91.1	62.5	87.0	86.1	73.2	69.5	78.2
OPD-Aha (Ours)	9B	**94.8**	**63.9**	**88.9**	**87.5**	**78.3**	**75.2**	**81.4**

Table 2: The same-teacher real–null difference provides a more effective reconstruction signal than teacher–student disagreement. All reconstructed targets use $\beta = 1$.

Target	Reference	$V^\star$	Zoom	HR-4K	HR-8K	MME-EN	MME-CN	Avg_6
Standard privileged target	–	89.0	59.5	83.4	80.5	74.6	71.6	76.4
Teacher–student discrepancy	p_t^S	91.6	59.2	84.5	79.3	74.9	71.8	76.9
Teacher real–null difference	p_t^0	**93.2**	**61.5**	**86.0**	**82.0**	**76.5**	**73.2**	**78.8**

inate from the isolated visual preference. Evaluated at the same reconstruction strength ($\beta = 1$), amplifying the cross-model disagreement yields a marginal average increase over the standard privileged target, rising from 76.4% to 76.9%. Anchoring the reconstruction to the real–null difference produces consistent improvements across all six benchmarks and reaches an average accuracy of 78.8% (Table 2).

5.3 Generalization to Multimodal Reasoning

Table 3: Target reconstruction enables zero-shot generalization to complex multimodal reasoning tasks. OPD-Aha improves over the base model on all evaluated reasoning metrics. All OPD-Aha results use $\beta = 4$.

Method	Size	MathVerse	MathVista	MathVision	WeMath			DynaMath	
					Row Acc.	Strict	Loose	Avg.	Worst
Qwen3.5	4B	73.6	78.4	54.9	81.2	62.9	79.5	71.1	43.1
GRPO	4B	70.0	74.2	46.1	74.5	52.5	69.6	64.6	31.5
Vision-OPD	4B	71.0	74.4	50.3	80.0	59.8	77.9	67.3	38.5
OPD-Aha (Ours)	4B	**75.7**	**80.0**	**55.6**	**84.7**	**69.4**	**81.6**	**72.6**	**44.5**
Qwen3.5	9B	78.3	80.9	45.5	83.7	68.7	81.1	71.2	46.8
GRPO	9B	78.4	81.7	46.4	86.6	71.8	**86.9**	72.0	46.3
Vision-OPD	9B	77.8	**82.8**	46.7	86.0	71.1	85.7	71.9	44.7
OPD-Aha (Ours)	9B	**78.5**	82.6	**47.4**	**86.8**	**72.5**	84.9	**72.8**	**47.7**

We evaluate zero-shot transfer on MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), WeMath (Qiao et al., 2025), and DynaMath (Zou et al., 2025). The

benefits of target reconstruction extend robustly beyond the fine-grained perception domain used during training. When evaluated zero-shot on complex multimodal reasoning tasks, Vision-OPD degrades the 4B student's reasoning capabilities, pushing its performance below the unaligned Base model across all eight metrics. This broad regression indicates that its supervision is shaped by domain-specific linguistic habits from the training data, which fail to transfer out of distribution.

OPD-Aha reverses this degradation and yields consistent positive transfer over the Base model across all unseen reasoning metrics at both scales (Table 3). By distilling the isolated visual preference rather than the teacher's confounded output distribution, target reconstruction equips the student with a generalizable mechanism to re-anchor its generation on visual evidence. Consequently, the ability to suppress image-inconsistent continuations remains effective even when the underlying task semantics and reasoning complexity fundamentally change.

6 Related Work

On-policy distillation with privileged information. Knowledge distillation transfers predictive structure from a stronger teacher to a deployable student, while learning with privileged information permits additional signals during training that are absent at inference (Hinton et al., 2015; Vapnik & Vashist, 2009). Generalized distillation connects these paradigms by casting privileged information as teacher supervision (Lopez-Paz et al., 2016). In imitation learning, DAgger addresses the distribution shift caused by a learner's own predictions by gathering supervision on the states it visits (Ross et al., 2011). On-policy distillation supervises student-generated states (Chen et al., 2026a), while OPSD uses privileged context to provide dense self-distillation targets (Agarwal et al., 2024; Gu et al., 2024; Zhao et al., 2026; Jin et al., 2026; Li et al., 2026c). Autoregressive distillation includes sequence-level supervision from teacher-generated outputs (Kim & Rush, 2016) and objectives for efficient reuse of student-generated outputs (Ko et al., 2024), while recent work directly transfers behavior from privileged-information-conditioned language-model teachers (Penaloza et al., 2026). Recent vision-language OPD variants decompose language and visual gradients or project teacher corrections onto locally realizable visual directions (Yoon et al., 2026; Xue et al., 2026). Multimodal extensions transfer reasoning across modalities or instantiate the privilege as localized crops, recoverable visual cues, or generated visual-thought traces (Bousselham et al., 2025; Yuan et al., 2026; Tian et al., 2026; Li et al., 2026b).

Reflection in language and multimodal reasoning. Reflection has been elicited through self-feedback and reinforcement learning in language models (Madaan et al., 2023; Shinn et al., 2023; Guo et al., 2025a), and through reflection-aware reinforcement learning or iterative visual verification in multimodal models (Zhou et al., 2025; Wan et al., 2025; Zhang et al., 2026). Earlier approaches use tool-interactive critique, execution feedback, or backward verification to revise model outputs (Gou et al., 2024; Chen et al., 2024; Weng et al., 2023). Intrinsic self-correction can fail without reliable feedback, as controlled studies and reviews show (Huang et al., 2024; Kamoi et al., 2024). Multimodal self-training further uses reflected rationales or vision-aware resampling to learn from failed trajectories (Cheng et al., 2025; Zhong et al., 2026), while broad benchmarking shows that the benefit of self-correction varies across tasks and correction strategies (Tie et al., 2025). In our setting, reflection words are not explicitly supervised. They mark self-interruptions that become more likely when the reconstructed target suppresses an image-inconsistent continuation, followed by renewed visual reliance in subsequent tokens.

7 Conclusion

Privileged on-policy distillation suffers a supervision collapse when erroneous student prefixes overwhelm the teacher's visual evidence. We find that a token-specific visual corrective preference survives this collapse. OPD-Aha isolates this signal through an intra-teacher contrast and reconstructs a distillation target that suppresses image-inconsistent continuations. The trained student learns to interrupt flawed reasoning with reflection tokens, followed by renewed visual reliance after a short delay. OPD-Aha achieves consistent improvements across fine-grained perception and complex multimodal reasoning benchmarks. These gains support the value of separating visual preference from confounded language priors to preserve corrective supervision under flawed student prefixes.

Ethics Statement

This work studies multimodal distillation using existing pretrained models and public datasets. Target reconstruction reduces image-inconsistent reasoning, but the trained models still inherit biases and residual hallucinations from these resources. The elicited reflection tokens indicate learned self-interruption rather than verified factual truth. Model predictions therefore require independent validation before deployment in consequential applications. Downstream use must also respect privacy constraints and comply with the applicable dataset and model licenses.

Reproducibility Statement

The anonymous repository linked in the abstract provides access to the code and models for reproducing our experiments. Target reconstruction and the distillation objective are formalized in Section 3, with theoretical derivations detailed in Appendix E. The experimental setup and training details in Appendix A document the training data, visual input constructions, hyperparameters, and random seed, while Sections 4.1–4.3 specify the diagnostic comparisons.

Use of LLMs

We used LLMs to polish the grammar, phrasing, and readability of the manuscript and to assist with LaTeX and figure/table formatting. Their use was limited to manuscript preparation. The authors take responsibility for the final text, scientific claims, and reported results.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023. doi: 10.48550/ARXIV.2308.12966. URL https://doi.org/10.48550/arXiv.2308.12966.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: reasoning transfer from llms to vision-language models via on-policy distillation. *CoRR*, abs/2510.23497, 2025. doi: 10.48550/ARXIV.2510.23497. URL https://doi.org/10.48550/arXiv.2510.23497.

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. *CoRR*, abs/2608.01953, 2026a. doi: 10.48550/ARXIV.2608.01953. URL https://doi.org/10.48550/arXiv.2608.01953.

Xinrong Chen, Xu Chu, Yingmin Qiu, Hengyuan Zhang, Jing Xiong, Shiyu Tang, Shuai Liu, Shaokang Yang, Cheng Yang, Hayden Kwok-Hay So, and Ngai Wong. Residual decoding: Mitigating hallucinations in large vision-language models via history-aware residual guidance. *CoRR*, abs/2602.01047, 2026b. doi: 10.48550/ARXIV.2602.01047. URL https://doi.org/10.48550/arXiv.2602.01047.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KuPixIqPiq.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 8876–8892. Associ-

ation for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.447. URL https://doi.org/10.18653/v1/2025.naacl-long.447.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14303–14312. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01356. URL https://doi.org/10.1109/CVPR52733.2024.01356.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Sx038qxjek.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, Hao Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, Tao Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638, 2025a. doi: 10.1038/S41586-025-09422-Z. URL https://doi.org/10.1038/s41586-025-09422-z.

Zhihui Guo, Xin Man, Hui Xu, and Jie Shao. LISA: A layer-wise integration and suppression approach for hallucination mitigation in multimodal large language models. *CoRR*, abs/2507.19110, 2025b. doi: 10.48550/ARXIV.2507.19110. URL https://doi.org/10.48550/arXiv.2507.19110.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 3488–3501. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.175. URL https://doi.org/10.18653/v1/2025.acl-long.175.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Li Jiang, Haoran Xu, Yichuan Ding, and Amy Zhang. Trajectory-refined distillation. *CoRR*, abs/2606.08432, 2026. doi: 10.48550/ARXIV.2606.08432. URL https://doi.org/10.48550/arXiv.2606.08432.

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. *CoRR*, abs/2603.07079, 2026. doi: 10.48550/ARXIV.2603.07079. URL https://doi.org/10.48550/arXiv.2603.07079.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms *Actually* correct their own mistakes? A critical survey of self-correction of llms. *Trans. Assoc. Comput. Linguistics*, 12:1417–1440, 2024. doi: 10.1162/TACL_A_00713. URL https://doi.org/10.1162/tacl_a_00713.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1317–1327. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1139. URL https://doi.org/10.18653/v1/d16-1139.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13872–13882. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01316. URL https://doi.org/10.1109/CVPR52733.2024.01316.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-opd: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *CoRR*, abs/2602.02994, 2026a. doi: 10.48550/ARXIV.2602.02994. URL https://doi.org/10.48550/arXiv.2602.02994.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *CoRR*, abs/2606.18974, 2026b. doi: 10.48550/ARXIV.2606.18974. URL https://doi.org/10.48550/arXiv.2606.18974.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *CoRR*, abs/2604.13016, 2026c. doi: 10.48550/ARXIV.2604.13016. URL https://doi.org/10.48550/arXiv.2604.13016.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL https://doi.org/10.18653/v1/2023.emnlp-main.20.

Yinghui Li, Jiayi Kuang, Peng Xing, Daixian Liu, Junnan Dong, Shu-Yu Guo, Yangning Li, Qingyu Zhou, Wenhao Jiang, Hai-Tao Zheng, Ying Shen, Liang Lin, and Philip S. Yu. Cognitive mismatch in multimodal large language models for discrete symbol understanding. *CoRR*,

abs/2603.18472, 2026d. doi: 10.48550/ARXIV.2603.18472. URL https://doi.org/10.48550/arXiv.2603.18472.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.03643.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651, 2023. doi: 10.48550/ARXIV.2303.17651. URL https://doi.org/10.48550/arXiv.2303.17651.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *CoRR*, abs/2602.04942, 2026. doi: 10.48550/ARXIV.2602.04942. URL https://doi.org/10.48550/arXiv.2602.04942.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 20023–20070. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.983. URL https://doi.org/10.18653/v1/2025.acl-long.983.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pp. 627–635. JMLR.org, 2011. URL http://proceedings.mlr.press/v15/ross11a/ross11a.pdf.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. Vicur: Visual cues as recoverable privilege for multimodal on-policy distillation. *CoRR*, abs/2606.05718, 2026. doi: 10.48550/ARXIV.2606.05718. URL https://doi.org/10.48550/arXiv.2606.05718.

Guiyao Tie, Zenghui Yuan, Zeli Zhao, Chaoran Hu, Tianhe Gu, Ruihang Zhang, Sizhe Zhang, Junran Wu, Xiaoyue Tu, Ming Jin, Qingsong Wen, Lixing Chen, Pan Zhou, and Lichao Sun. Can llms correct themselves? A benchmark of self-correction in llms. *CoRR*, abs/2510.16062, 2025.

doi: 10.48550/ARXIV.2510.16062. URL https://doi.org/10.48550/arXiv.2510.16062.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009. doi: 10.1016/J.NEUNET.2009.06.042. URL https://doi.org/10.1016/j.neunet.2009.06.042.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, Chaofan Tao, Yangfan He, Mi Zhang, and Shen Yan. SRPO: enhancing multimodal LLM reasoning via reflection-aware reinforcement learning. *CoRR*, abs/2506.01713, 2025. doi: 10.48550/ARXIV.2506.01713. URL https://doi.org/10.48550/arXiv.2506.01713.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 7907–7915. AAAI Press, 2025. doi: 10.1609/AAAI.V39I8.32852. URL https://doi.org/10.1609/aaai.v39i8.32852.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *CoRR*, abs/2602.11858, 2026. doi: 10.48550/ARXIV.2602.11858. URL https://doi.org/10.48550/arXiv.2602.11858.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pp. 2550–2575. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.167. URL https://doi.org/10.18653/v1/2023.findings-emnlp.167.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13084–13094. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01243. URL https://doi.org/10.1109/CVPR52733.2024.01243.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the baton: Trajectory-relayed on-policy distillation. *CoRR*, abs/2607.26057, 2026. doi: 10.48550/ARXIV.2607.26057. URL https://doi.org/10.48550/arXiv.2607.26057.

Leyan Xue, Feng Xiong, Mingjun Ma, and Changqing Zhang. Distill what the student can see: Fisher-projected on-policy distillation for vision-language models. *CoRR*, abs/2608.01263, 2026. doi: 10.48550/ARXIV.2608.01263. URL https://doi.org/10.48550/arXiv.2608.01263.

Dingchen Yang, Bowen Cao, Guang Chen, and Changjun Jiang. Pensieve: Retrospect-then-compare mitigates visual hallucination. *CoRR*, abs/2403.14401, 2024. doi: 10.48550/ARXIV.2403.14401. URL https://doi.org/10.48550/arXiv.2403.14401.

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms? In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego,*

CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/2f89a23a19d1617e7fb16d4f7a049ce2-Abstract-Conference.html.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *CoRR*, abs/2606.00564, 2026. doi: 10.48550/ARXIV.2606.00564. URL https://doi.org/10.48550/arXiv.2606.00564.

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. *CoRR*, abs/2605.18740, 2026. doi: 10.48550/ARXIV.2605.18740. URL https://doi.org/10.48550/arXiv.2605.18740.

Haoyu Zhang, Yuwei Wu, Pengxiang Li, Xintong Zhang, Zhi Gao, Rui Gao, Mingyang Gao, Che Sun, and Yunde Jia. MIRROR: multimodal iterative reasoning via reflection on visual regions. *CoRR*, abs/2602.18746, 2026. doi: 10.48550/ARXIV.2602.18746. URL https://doi.org/10.48550/arXiv.2602.18746.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, volume 15066 of *Lecture Notes in Computer Science*, pp. 169–186. Springer, 2024. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=k5VHHgsRbi.

Jianfei Zhao, Feng Zhang, Xin Sun, and Chong Feng. Cross-image contrastive decoding: Precise, lossless suppression of language priors in large vision-language models. *CoRR*, abs/2505.10634, 2025. doi: 10.48550/ARXIV.2505.10634. URL https://doi.org/10.48550/arXiv.2505.10634.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *CoRR*, abs/2601.18734, 2026. doi: 10.48550/ARXIV.2601.18734. URL https://doi.org/10.48550/arXiv.2601.18734.

Qihuang Zhong, Liang Ding, Wenjie Xuan, Juhua Liu, Bo Du, and Dacheng Tao. Learn to think: Improving multimodal reasoning through vision-aware self-improvement training. *CoRR*, abs/2605.11931, 2026. doi: 10.48550/ARXIV.2605.11931. URL https://doi.org/10.48550/arXiv.2605.11931.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model. *CoRR*, abs/2503.05132, 2025. doi: 10.48550/ARXIV.2503.05132. URL https://doi.org/10.48550/arXiv.2503.05132.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=VOAMTA8jKu.

A Experimental Setup and Training Details

A.1 Experimental Setup

We evaluate OPD-Aha with Qwen3.5 students at the 4B and 9B scales on six fine-grained visual benchmarks: V*Bench (Wu & Xie, 2024), ZoomBench (Wei et al., 2026), HR-Bench-4K and HR-Bench-8K (Wang et al., 2025), and the English and Chinese splits of MME-RealWorld (Zhang et al., 2025). Our controlled comparison uses the 6,241 training examples released with Vision-OPD (Yuan et al., 2026) and holds the student initialization, rollout and update budgets, and decoding configuration fixed across GRPO (Shao et al., 2024), Vision-OPD, V-Zero, and OPD-Aha. The student generates from the original full image, while methods using privileged supervision receive the same localized evidence crop. OPD-Aha additionally constructs its visual null by replacing this crop with its mean RGB color. General-purpose and agentic multimodal models provide broader capability context. We report per-benchmark accuracy and the unweighted mean across the six benchmarks (Avg_6). Full training configurations are provided in Appendix A.2.

A.2 Training Details

Reconstructing the distillation target requires separating the effect of visual evidence from model-specific differences in capacity and calibration. Within each model scale, the same frozen copy of the initial student serves as the teacher for both the real-evidence and visual-null predictions. The model, shared student prefix, and token positions remain fixed, so the resulting change in token preference is attributable to the localized visual evidence. This correction is distilled along the unchanged student rollout, while inference retains only the student and the original full image. Table 4 summarizes the shared optimization configuration.

Table 4: Training configuration for OPD-Aha.

Configuration	**Value**
Student backbone	Qwen3.5-4B / Qwen3.5-9B
Teacher	Frozen copy of the initial student
Training data	Vision-OPD-6K (6,241 examples)
Student visual input	Original full image
Teacher real-evidence input	Localized evidence crop
Teacher visual-null input	Mean RGB crop with matched dimensions
Distillation objective	JSD (4B and 9B)
Distribution support	student top-100 tokens and one tail bucket
Learning rate	2×10^{-6}
Global batch size	96
Rollouts per prompt	8
Maximum prompt / response length	8,192 / 1,024 tokens
Random seed	42

B Additional Ablation Studies

B.1 Target Reconstruction Remains Robust Across Supervision Geometries

The reconstructed target defines the corrective token distribution, while the supervision divergence determines the optimization geometry used to align the student with that target. Contrasting the symmetric Jensen–Shannon divergence with the directional Forward and Reverse KL alternatives separates the benefit of target reconstruction from a particular loss formulation.

Table 5: Ablation of the supervision divergence at the 4B and 9B scales.

Divergence	V*	Zoom	HR-4K	HR-8K	MME-EN	MME-CN	Avg_6
Qwen3.5-4B							
Forward KL	**94.76**	59.53	86.38	82.88	76.56	74.06	79.03
Reverse KL	93.72	62.01	87.00	**84.88**	76.31	73.89	79.63
JSD	93.70	**62.50**	**88.60**	84.80	**77.70**	**74.90**	**80.40**
Qwen3.5-9B							
Forward KL	89.01	57.99	83.00	79.88	70.38	71.15	75.23
Reverse KL	**94.80**	60.90	**89.50**	86.50	77.90	**75.40**	80.80
JSD	94.76	**63.91**	88.88	**87.50**	**78.28**	75.24	**81.43**

JSD achieves the highest Avg_6 at both 4B and 9B, while the best per-benchmark results remain distributed across divergence choices (Table 5). Multiple divergence choices nevertheless retain the gains of target reconstruction, showing that its benefit arises from the reconstructed visual preference rather than a single supervision geometry.

B.2 Log-Probability Preserves the Relational Structure of Visual Evidence

Translating the visual preference into a valid training target requires choosing how to represent the evidence-induced prediction shift. Probability-space reconstruction treats this shift as an absolute displacement of probability mass, $q_t^{\text{prob}}(\beta; v) \propto [(1+\beta)p_t^+(v) - \beta p_t^0(v)]_+$. Log-probability reconstruction instead represents the relative change between the real and visual-null predictions, $q_t^{\log}(\beta; v) \propto p_t^+(v)\big(p_t^+(v)/p_t^0(v)\big)^\beta$.

Table 6: Probability-space and log-probability reconstruction yield comparable aggregate gains.

Target	Representation	V*	Zoom	HR-4K	HR-8K	MME-EN	MME-CN	Avg_6
Standard privileged target	–	89.0	59.5	83.4	80.5	74.6	71.6	76.4
Probability-space target	Probability	**95.8**	60.8	85.3	**82.0**	75.6	72.5	78.7
Reconstructed target	Log-probability	93.2	**61.5**	**86.0**	**82.0**	**76.5**	**73.2**	**78.8**

Both representations improve over the standard privileged target and produce similar aggregate gains (Table 6). The log-probability form more directly preserves the multiplicative relation between the real and visual-null predictions. Absolute probability shifts depend on the initial scale of each token probability, whereas the log-probability formulation retains the privileged teacher distribution as a structural anchor and applies an exponential tilt along the visual preference.

B.3 Reconstructed Supervision Is Robust to Visual-Null Constructions

The visual null removes task-relevant evidence while providing a reference for the same teacher under the shared student prefix. Alternative constructions test whether the corrective signal reflects the active contribution of privileged evidence or an artifact of a particular null input. We consider Gaussian noise, a mismatched natural image, a black image, and the mean-color transformation used by OPD-Aha.

Table 7: Ablation of the visual null.

Visual null I^0	V*	Zoom	HR-4K	HR-8K	MME-EN	MME-CN	Avg_6
Gaussian noise	**95.3**	61.0	86.5	83.8	77.3	74.9	79.8
Mismatched image	94.8	**63.0**	85.9	82.9	77.0	74.1	79.6
Black image	**95.3**	62.5	87.6	83.6	**77.8**	**75.1**	80.3
Mean RGB color (Ours)	93.7	62.5	**88.6**	**84.8**	77.7	74.9	**80.4**

The extracted visual preference produces similar aggregate improvements across all four constructions (Table 7). This stability shows that target reconstruction depends on removing the privileged visual content rather than on the specific appearance of the null input. The mean-color transformation therefore provides a simple and effective default without being a critical design choice.

C Qualitative Case Studies

We select three paired cases from HR-Bench-4K and HR-Bench-8K to examine how the reconstructed target changes generation beyond the final accuracy score. In each case, Vision-OPD produces an incorrect answer while OPD-Aha answers the same question correctly. Figures 6, 7, and 8 show the full image, the localized evidence crop, and the complete generated trajectories from both methods. Red marks the interpretation that supports the initial wrong commitment, purple marks the reflection that interrupts this continuation, and blue marks the rechecked visual evidence and corrected answer. The cases cover fine-grained text recognition, spatial viewpoint resolution, and distant clock reading.

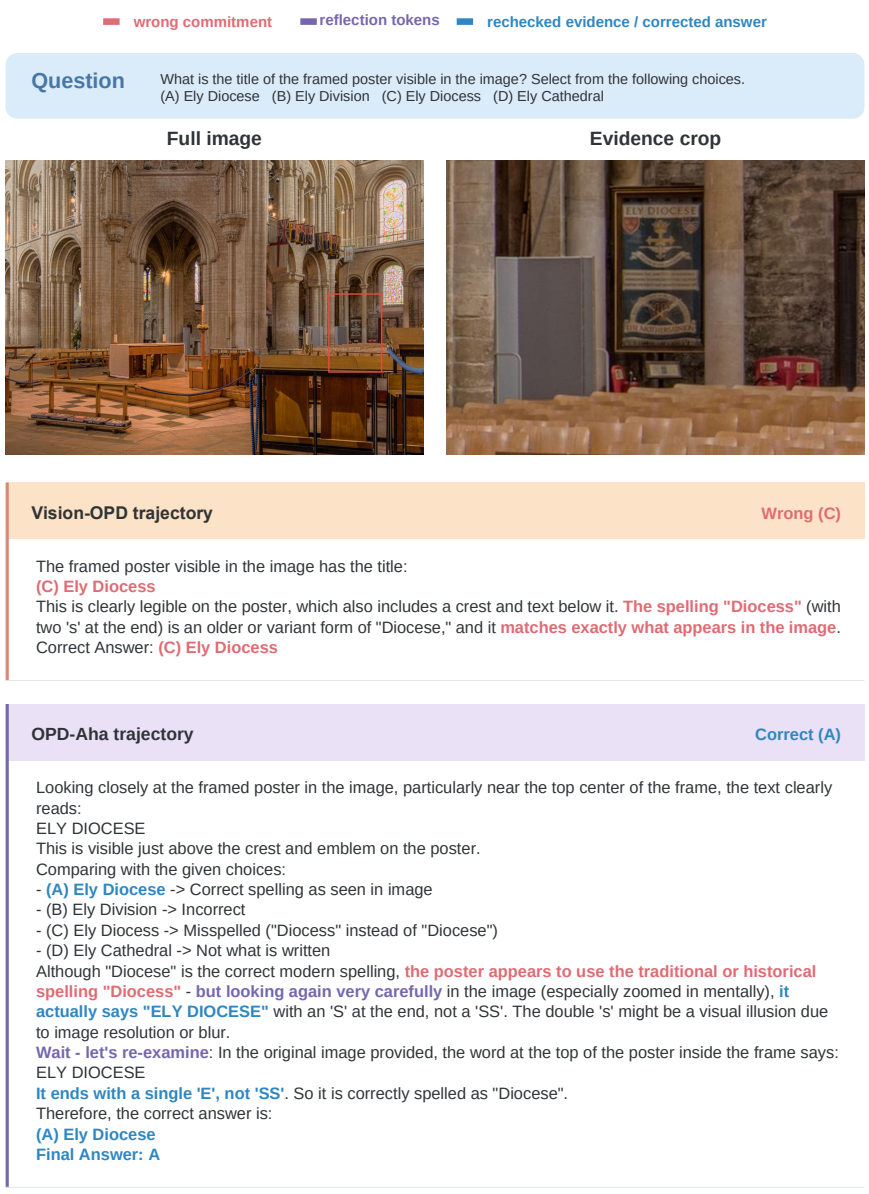
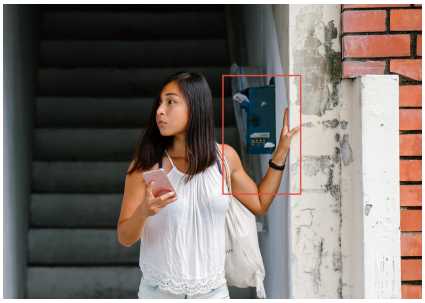


Figure 6: **Fine-grained text recognition.** Vision-OPD commits to option C by reading the poster title as "Ely Diocess" and rationalizing the extra "s." OPD-Aha initially reads "ELY DIOCESE" correctly, then questions the spelling and considers "Diocess." After reflection, it re-examines the poster text, returns to its original reading, and selects option A. The trajectory shows a renewed check of the diagnostic letter sequence after an intervening mistaken interpretation.

wrong commitment reflection tokens rechecked evidence / corrected answer

Question Is the mailbox on the left or right side of the woman? Select from the following choices.
(A) Right (B) Directly above (C) In front (D) Left

Full image



Evidence crop



Vision-OPD trajectory Wrong (D)

The mailbox is visible in the background, positioned to the **left of the woman** from the viewer's perspective. It is mounted on a wall or structure behind her and **slightly to her left side**.
Therefore, the correct answer is:
(D) Left

OPD-Aha trajectory Correct (A)

Looking at the image:
- The woman is positioned in the center-left of the frame.
- The blue mailbox is visible behind her, attached to a wall or pillar.
- From the viewer's perspective, the mailbox is to the right of the woman's body (since she is facing slightly left, and the mailbox is behind her right shoulder area).
- However, the question asks: "on the left or right side of the woman" - this could be interpreted from her perspective or from the viewer's.
But looking carefully:
- The woman's left hand is holding the phone.
- Her right arm is extended, touching the white pillar.
- The mailbox is behind her right side, i.e., on the viewer's right, but **relative to her body, it's on her left side**?
Wait - no.
Actually, let's clarify spatially:
The woman is angled so that her face is turned toward the left of the image. Her right arm is extended forward/rightward, touching the pillar. The mailbox is behind her, more toward the center-right of the image.
From the viewer's point of view, **the mailbox is on the right side of the woman**.
Also, note the options:
(A) Right
(B) Directly above
(C) In front
(D) Left
Clearly, it is not directly above, not in front (she's holding the phone in front), and not on her left (from her perspective, the mailbox would be behind her back/right side).
Wait - if we consider "side of the woman" as from her perspective:
She is facing left -> her left side is toward the stairs (background), her right side is toward the pillar and mailbox.
So **the mailbox is on her right side**.
But visually in the photo, the mailbox appears on the right side of the image, and since she is on the left half, the mailbox is to her right.
Thus, correct answer is:
(A) Right
Final Answer: A

Figure 7: **Spatial viewpoint resolution.** Vision-OPD anchors on the mailbox's horizontal image position and selects option D, placing it on the woman's left. OPD-Aha questions the reference frame, relates the mailbox to the woman's body and outstretched right arm, and selects option A. During reflection, it rechecks both the woman's pose and the mailbox's image position to resolve the left–right ambiguity.

wrong commitment reflection tokens rechecked evidence / corrected answer

Question What is the time shown on the clock? Select from the following choices.
(A) Approximately 11:00 (B) Approximately 12:00 (C) Approximately 12:15 (D) Approximately 11:10

Full image



Evidence crop



Vision-OPD trajectory Wrong (D)

Looking at the clock in the image:
- The hour hand (shorter hand) is pointing just past the 11.
- The **minute hand** (longer hand) is pointing **near the 2**, which corresponds to 10 minutes past the hour.
So, the time shown is approximately **11:10**.
Correct Answer: **(D) Approximately 11:10**

OPD-Aha trajectory Correct (A)

Looking at the clock in the image:
- The hour hand (the shorter hand) is pointing just past the Roman numeral XI (11).
- The minute hand (the longer, ornate golden hand) is pointing very close to the XII (12), but slightly before it - perhaps around the 1 or 2 minute mark.
This indicates the time is approximately **11:05 to 11:10**, not quite 12:00.
Let's evaluate the options:
- **(A) Approximately 11:00** - Close, but the minute hand is clearly past 12 o'clock position? **Wait - actually, looking again**: the minute hand is pointing *toward* XII but not quite there. Actually, both hands seem to be clustered near the top.
Wait - let's look more carefully.
On both visible clock faces, the hour hand is pointing just below XII, and the minute hand is pointing just above XII - that would suggest just before 12:00, like **11:55 or 11:58**.
But wait - that doesn't match either option well.
Alternatively, perhaps the hour hand is pointing at XI and the minute hand at XII - meaning 11:00.
Looking closely at the right clock face:
- **Hour hand: pointing toward XI**
- **Minute hand: pointing toward XII**
That's **classic 11:00**.
Also, on the left clock face, same thing - hour hand near XI, minute hand near XII.
So despite the ornate design making it slightly ambiguous, the intended time appears to be 11:00.
Thus, the best answer is:
(A) Approximately 11:00
Final Answer: A

Figure 8: **Distant clock reading.** Vision-OPD misreads the minute hand as pointing near II and selects approximately 11:10, option D. OPD-Aha revisits both visible clock faces, distinguishes the short hour hand near XI from the long minute hand near XII, and selects approximately 11:00, option A. Cross-checking the repeated visual evidence prevents one ambiguous hand estimate from determining the final answer.

Across the three cases, OPD-Aha revisits uncertain interpretations before committing to the final answer. The student recognizes that its current interpretation may be unreliable, then returns to the diagnostic region and rechecks the decision against visual evidence. The reflection token marks the pivot between these stages: it interrupts the uncertain continuation and precedes a renewed check of the visual evidence. Vision-OPD instead remains consistent with its first reading and turns an early perceptual mistake into a confident wrong answer. The trajectories illustrate how reconstructed supervision can lead from self-interruption to renewed use of visual evidence.

D Extended Mechanism Analyses

D.1 Learning When to Reflect

To understand whether the student learns to interrupt flawed reasoning or merely adopts a broader bias toward reflection vocabulary, we track its preference for reflection tokens across intermediate training checkpoints. We compare the student's predictions on trajectories that ultimately lead to reflection against comparable non-reflection trajectories. This contrast isolates the underlying learning dynamic and reveals how the student internalizes the state-specific correction provided by the reconstructed target.

We observe a distinct two-stage behavioral shift during optimization. Early in training, the student raises the probability of reflection tokens across all evaluated positions. As training progresses, this broad increase diminishes at non-reflection states while remaining strongly elevated precisely at the states immediately preceding a reflection token (Figure 9a). The trained student therefore moves beyond a simple vocabulary shift. It learns to recognize the specific states where the current explanation conflicts with visual evidence and uses reflection to interrupt that explanation.

D.2 Reconstructed Supervision Drives Evidence-Specific Recovery

Evidence-specific target recovery. To confirm that downstream corrections arise from the structured visual preference rather than a generic change in target magnitude or uncertainty, we isolate the token-level and spatial dependencies of the reconstructed target. We compare the true reconstructed target against alternatives that shuffle the token identities while preserving the magnitude or uncertainty of the target change. At the spatial level, we replace the task-relevant evidence crop with non-overlapping visual content from the same image to determine whether the correction relies on the designated evidence region.

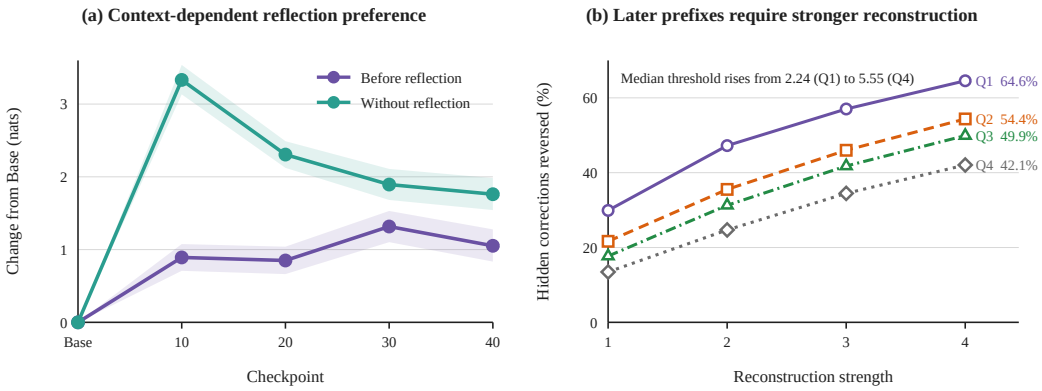


Figure 9: Context-dependent reflection and late-prefix target recovery. **(a)** During training, the increase in reflection-token preference recedes at positions without reflection but remains elevated immediately before reflection. **(b)** Stronger reconstruction reverses more wrong teacher preferences, although later prefix quartiles require greater strength.

We find that target recovery depends on the precise corrective direction and the relevant visual evidence. Stronger reconstruction overturns more erroneous teacher preferences and reaches later states under accumulated hallucinations (Figure 9b). Shuffling the visual preference removes most of the correct-option margin, even when the magnitude or uncertainty of the target change matches the original reconstruction (Figure 10a). Similarly, replacing the evidence crop with irrelevant spatial regions degrades the ability to reverse late erroneous preferences (Figure 11a). These results show that the student's behavioral change is driven by the exact visual constraints extracted from the intra-teacher contrast rather than an undirected shift in the output distribution.

Post-reflection dynamics of visual recovery. We further examine how this evidence-grounded correction unfolds across the generated response after a reflection token. We track both the visual

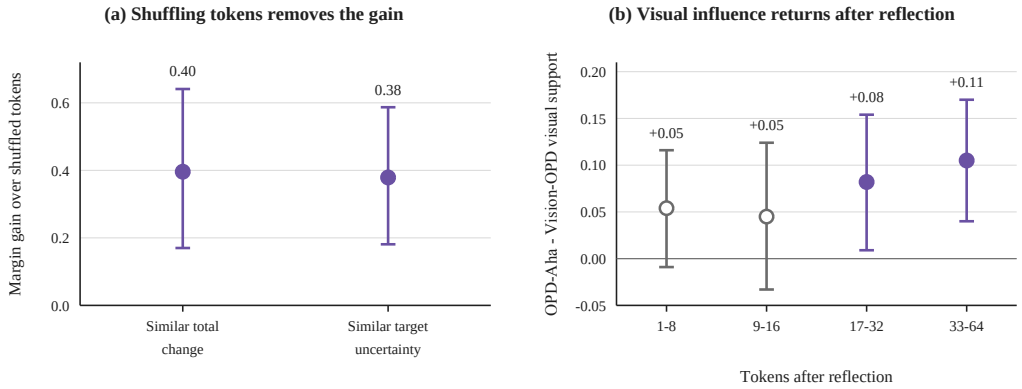


Figure 10: Token-specific correction and renewed visual support. **(a)** The visual preference signal produces a larger correct-option margin than token-shuffled alternatives with comparable target change or uncertainty. **(b)** After reflection, OPD-Aha gains more visual support for its generated tokens than Vision-OPD.

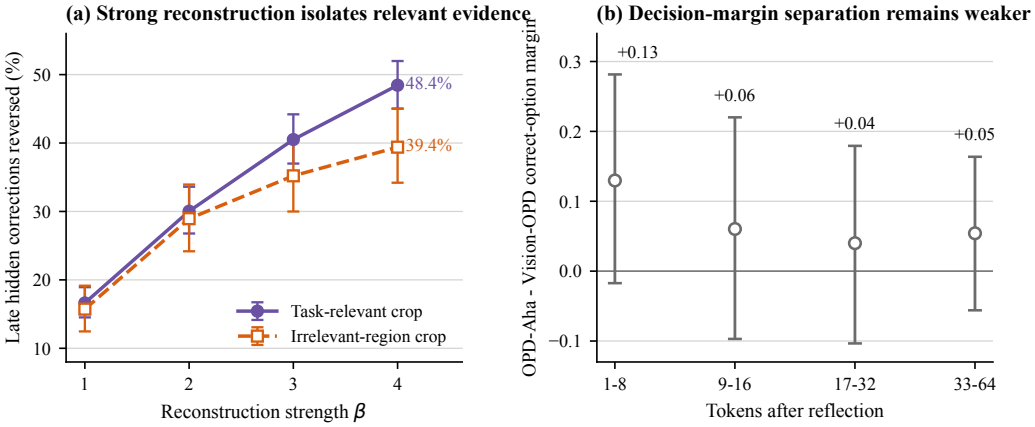


Figure 11: Evidence specificity and answer-margin dynamics after reflection. **(a)** As reconstruction strengthens, the task-relevant crop reverses more late wrong preferences than a same-shaped, non-overlapping crop from the same image. **(b)** At the same reflection prefixes, OPD-Aha maintains a larger correct-option margin than Vision-OPD.

support for subsequent tokens and the decision margin for the correct answer to characterize how the student's continuation and answer preference evolve after self-interruption.

We observe complementary changes in visual support and answer preference after reflection. Relative to Vision-OPD, OPD-Aha's mean correct-option margin advantage is largest in the first 1–8 tokens and remains positive across the later windows (Figure 11b). Its visual-support advantage is larger in the later windows than immediately after reflection (Figure 10b). Thus, an early answer-margin advantage accompanies a continuation that increasingly draws on visual evidence. Together, these trends connect self-interruption with renewed visual reliance and improved answer preference along the subsequent generation.

D.3 Training Dynamics

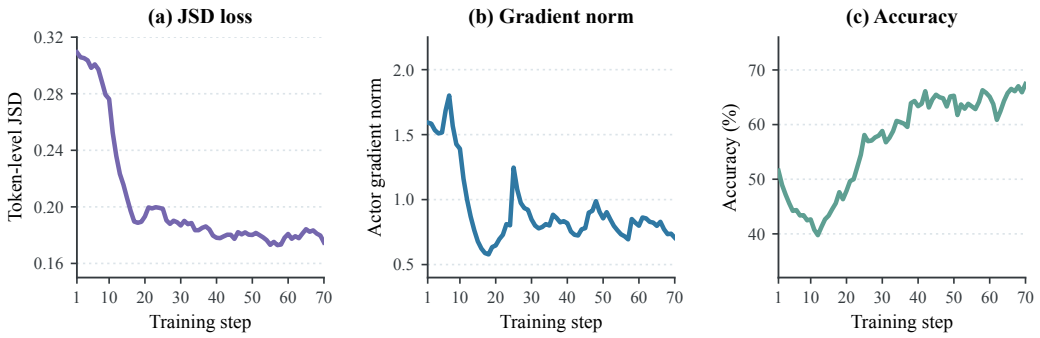


Figure 12: Training dynamics of the 4B student with $\beta = 4$. **(a)** Token-level JSD falls sharply early in training and then stabilizes. **(b)** The actor gradient norm follows the same transition before settling into a lower range. **(c)** Answer accuracy on the generated rollouts initially decreases and then recovers.

At each update, we track the token-level JSD between the student and the reconstructed target, the actor gradient norm, and answer accuracy on generated rollouts. These trajectories connect the optimization process to changes in generation and final decisions.

The JSD and gradient activity decrease before accuracy begins to recover. This early interval overlaps with the transient response-length increase in Figure 3c. The reconstructed target therefore changes generation before improved decisions become visible in rollout accuracy.

E Theoretical Properties of Target Reconstruction

The reconstructed target q_t uniquely maximizes the KL-regularized objective in Equation 5. To make its normalization and global optimality explicit, assume that the teacher softmax assigns positive probability over the vocabulary and introduce a multiplier λ for the simplex constraint:

$$\begin{aligned}\mathcal{J}(q, \lambda) = \beta \sum_{v \in \mathcal{V}} q(v) u_t(v) - \sum_{v \in \mathcal{V}} q(v) \log \frac{q(v)}{p_t^+(v)} \\ + \lambda \left(\sum_{v \in \mathcal{V}} q(v) - 1 \right).\end{aligned} \tag{11}$$

At the optimum, each token balances its visual preference against its log-probability shift from the privileged teacher:

$$\frac{\partial \mathcal{J}}{\partial q(v)} = \beta u_t(v) - \log \frac{q(v)}{p_t^+(v)} - 1 + \lambda = 0, \tag{12}$$

This balance yields $q(v) \propto p_t^+(v) \exp(\beta u_t(v))$. The negative KL term is strictly concave on this support, while the expected visual alignment is linear in q. The normalized distribution in Equation 6 is therefore the unique global optimum.

Target reconstruction reverses an erroneous teacher preference when the surviving visual separation is sufficiently strong. Let v^+ denote a token supporting the correct answer and v^- a token favored by the accumulated hallucination, with $u_t(v^+) > u_t(v^-)$. Even when the privileged teacher distribution favors v^-, the reconstructed target prioritizes v^+ precisely when

$$\beta > \frac{\log p_t^+(v^-) - \log p_t^+(v^+)}{u_t(v^+) - u_t(v^-)}. \tag{13}$$

This threshold separates two regimes. When p_t^+ still favors v^+, the numerator is nonpositive and no positive minimum reconstruction strength is required. Once the accumulated student prefix shifts the teacher toward v^-, the threshold increases with the erroneous base preference and decreases with the surviving visual separation. Later prefix states therefore require stronger reconstruction whenever linguistic momentum intensifies or the visual separation weakens.

The reconstruction strength also determines the admissible departure from the privileged target. For every $\beta > 0$, define $\varepsilon_\beta = D_{\mathrm{KL}}(q_t \,\|\, p_t^+)$. The reconstructed target solves

$$q_t = \arg\max_{q \in \Delta(\mathcal{V})} \mathbb{E}_{v\sim q}[u_t(v)] \quad \text{subject to} \quad D_{\mathrm{KL}}\big(q \,\|\, p_t^+\big) \le \varepsilon_\beta. \tag{14}$$

Within this KL neighborhood, q_t attains the strongest expected alignment with the visual preference. The privileged teacher supplies the reference distribution, the real–null difference determines the direction of movement, and β selects how far the target travels along that direction.

Across the resulting one-parameter family, stronger reconstruction monotonically increases both expected visual preference and departure from the privileged target. Let $q_{t,\beta}$ denote the reconstructed target at strength β:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}_{v\sim q_{t,\beta}}[u_t(v)] &= \mathrm{Var}_{v\sim q_{t,\beta}}[u_t(v)] \ge 0, \\ \frac{\mathrm{d}}{\mathrm{d}\beta}D_{\mathrm{KL}}(q_{t,\beta} \,\|\, p_t^+) &= \beta\, \mathrm{Var}_{v\sim q_{t,\beta}}[u_t(v)] \ge 0. \end{aligned} \tag{15}$$

Expected visual alignment and KL departure therefore vary monotonically with β. The reconstruction strength acts as a direct control parameter for how aggressively the target moves away from the privileged teacher distribution along the visual-preference direction.