PREVID: INTERVENTIONAL PREFIX REPLAY FOR VISUAL ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual distillation uses a teacher with better visual evidence to supervise responses sampled by a student model, but early visual mistakes remain in the response history and distort later supervision. We identify this missing pathway by showing that corrected visual evidence can change later predictions through the teacher's language context. Based on this finding, we introduce PReViD, which replays the same student response under controlled image and evidence contexts while preserving alignment with the original response. PReViD separates visual and contextual corrections, uses a placebo prefix with no visual evidence to filter generic text effects, and gives more weight to context changes that support the teacher's correction. Across benchmarks for detailed visual understanding, model sizes, and teacher settings, PReViD outperforms strong distillation baselines without changing inference. Code is available at `https://anonymous.4open.science/r/E02D2`.

1 INTRODUCTION

Multimodal large language models (MLLMs) now handle many visual understanding and reasoning tasks (Bai et al., 2025; Dong et al., 2025; Cheng et al., 2025). Their responses are autoregressive. Each prediction becomes context for the next. Recent studies show that text can outweigh conflicting visual evidence (Deng et al., 2025; Kancheti et al., 2026). An early visual error can then persist through later reasoning steps (Zhu et al., 2026b). Effective supervision must correct the error and the downstream state it creates.

This error propagation also affects knowledge distillation. Sequence-level distillation learns from teacher-generated responses (Kim & Rush, 2016). It may not cover states created by the student's own mistakes (Agarwal et al., 2024; Ko et al., 2025). Fixed offline distillation can also overfit errors in the teacher signal (Tiapkin et al., 2025). On-policy distillation (OPD) instead samples a student response. The teacher then scores that response (Agarwal et al., 2024; Zhu et al., 2026a). Visual OPD can strengthen this supervision by giving the teacher privileged visual information (Yuan et al., 2026; Wang et al., 2026; Li et al., 2026b). This information is unavailable to the student. Recent methods improve the teacher's view, select visually dependent states, or reconstruct evidence-sensitive targets (Liu et al., 2026; Zhang et al., 2026; Aniri et al., 2026). These methods preserve token-level alignment. They still score each token under the student-generated history.

However, the same alignment creates a blind spot when the history contains incorrect visual evidence, as shown in Figure 1. The teacher can correct an early statement at its own position. Later targets still condition on that statement. They do not show how the corrected statement would change later predictions. A stronger teacher does not solve this problem because it is queried under the same incorrect context. Generating a new response from privileged visual information would provide a corrected context, but the new response would not align token by token with the student's rollout. The key challenge is to expose the downstream effect of corrected evidence while retaining aligned supervision.

To measure this missing effect, we separate two paths by which privileged visual evidence changes a later prediction. First, we improve the visual input but leave its description in the response unchanged. This step isolates the change caused by the image. Next, we add an evidence prefix from the clear view while keeping the improved image and the complete student history fixed. The additional change

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

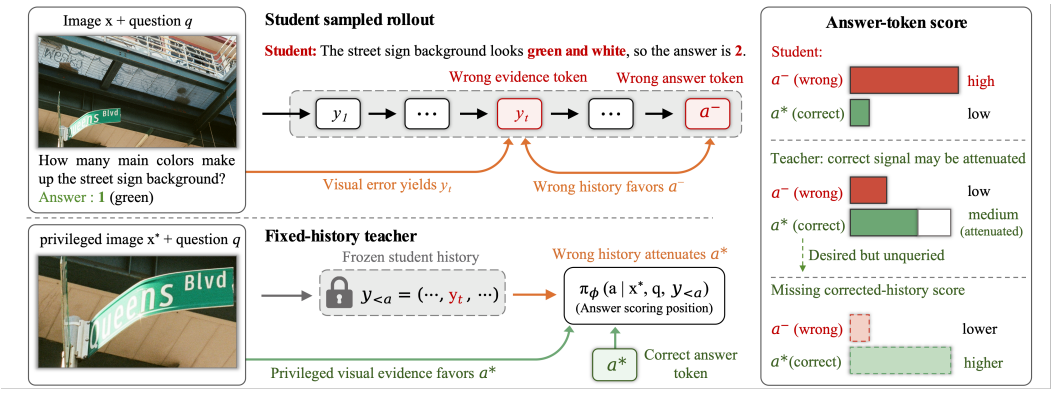


Figure 1: **Fixed-history visual OPD keeps incorrect evidence in the context for later supervision.** The student first states an incorrect visual observation and then gives the wrong answer. The teacher receives a privileged image but still scores the answer under the student's original history. The incorrect observation in that history can weaken the corrective signal. In this example, the privileged image is an evidence-focused crop. More generally, privileged visual evidence can come from stronger teacher perception of the original image.

measures how this prefix affects later predictions. The original history is never edited, so token-level alignment remains intact.

Figure 2 shows two findings. The evidence-prefix path accounts for 76% of the mean correction. This share rises to 81% for visually sensitive examples. A clearer image alone therefore misses much of the benefit from privileged visual information. The evidence-prefix effect also varies across examples. It can support the correct answer, have little effect, or favor a wrong answer. A uniform target would mix grounded corrections with generic effects from added text. This variation calls for selective transfer.

PReViD turns this analysis into a training objective for visual OPD. It performs the ordered comparison within the teacher's replay context. A matched placebo prefix estimates the generic effect of adding text. PReViD retains contextual changes that support the teacher's visual correction. All auxiliary prefixes are used only during training. Inference uses the original student model and image input.

The main contributions of this work are as follows.

- **We reveal a missing supervision pathway in fixed-history visual OPD.** These methods can correct an early visual error, but later targets still condition on that error. Our decomposition measures the omitted effect of adding teacher-side evidence before later predictions while preserving alignment with the student's response.
- **We introduce PReViD to recover this pathway without replacing the student rollout.** PReViD separates two sources of correction: the visual input and the language context. A matched placebo prefix filters generic effects of adding text. All auxiliary prefixes are used only during training.
- **We establish teacher-side evidence context as a measurable and selective source of supervision.** The full-corpus analysis quantifies its magnitude and variation across examples. Experiments across scales, teacher settings, and visual benchmarks test its selective use for training.

2 Related Work

2.1 On-policy distillation

Distillation for autoregressive models depends on who generates the training trajectory. Sequence-level distillation (Kim & Rush, 2016) trains on teacher responses. GKD (Agarwal et al., 2024) instead

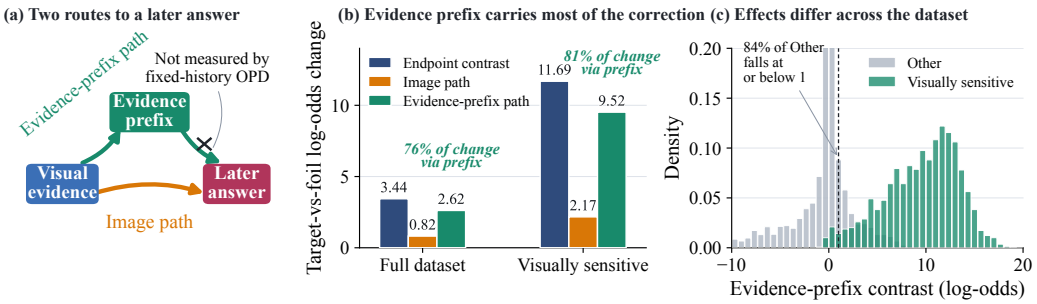


Figure 2: **Controlled image and evidence-prefix interventions change how the teacher supervises later answers.** (a) Visual evidence can affect a later answer through the image or through an evidence prefix. Fixed-history visual OPD measures the image path but not the evidence-prefix path. (b) On Vision-OPD-6K (Yuan et al., 2026), the evidence-prefix path accounts for 76% of the mean correction across all examples. The share is 81% for visually sensitive examples. These examples are correct under $p_t^{+,+}$ and wrong under $p_t^{-,-}$, so the subset depends on both endpoint contexts. (c) Evidence-prefix effects vary across examples. This variation motivates selective transfer.

queries the teacher on student rollouts. This reduces the mismatch between training states and the student's inference distribution. DistiLLM (Ko et al., 2024) adds adaptive replay. DistiLLM-2 (Ko et al., 2025) contrasts teacher and student responses. Speculative Knowledge Distillation (Xu et al., 2025) lets the teacher replace weak student proposals during sampling.

Recent work refines the objective and the supervision budget. HPD (Zhu et al., 2026a) mixes forward and reverse KL with lightweight approximate on-policy sampling. π-Distill and OPSD (Penaloza et al., 2026) transfer training-only privileged information to an unconditioned policy. TIP (Xu et al., 2026) selects tokens using student uncertainty and teacher disagreement. TOPD (Jiang & Ferraro, 2026) supplies near-future guidance. These methods improve coverage, efficiency, or token selection. Standard token-level OPD still evaluates the teacher under each realized student prefix. A wrong premise in that prefix therefore remains part of every later target.

2.2 Visual On-Policy Distillation

Earlier MLLM distillation mainly uses fixed traces or internal alignment. LLaVA-MoD (Shu et al., 2025) transfers output distributions and preferences. CompoDistill (Kim et al., 2026) aligns visual attention. Align-TI (Chen et al., 2026) aligns visual and response-token interactions. Masking-KD (Yu et al., 2026) hides salient student-prefix tokens to reduce reliance on textual cues. These methods strengthen visual learning, but they do not supervise states produced by an evolving student.

Visual OPD instead scores student rollouts with multimodal teacher signals. Many methods create asymmetry through the visual input. Vision-OPD (Yuan et al., 2026) uses evidence-focused crops. NOPD (Wang et al., 2026) pairs clean and corrupted views. S^2VOPD (Li et al., 2026b) creates the asymmetry through image augmentation. ViCuR (Tian et al., 2026) uses visual cues that the student can recover. Video-OPD (Li et al., 2026a) extends OPD to temporal grounding. V-Zero (Sun et al., 2026) gates trajectories with positive and negative visual views.

Recent methods also refine where and what to distill. VA-OPD (Liu et al., 2026) finds tokens that depend on visual detail. OPD-V (Aniri et al., 2026) selects tokens with balanced visual and textual support. PCD (Xiong et al., 2026) detects correctable perception failures. VCSD (Liang et al., 2026) contrasts the image with a content-erased control. VAD (Zhang et al., 2026) reconstructs the visually attributable correction. VGS (Yoon et al., 2026) steers gradients toward visual grounding. FP-OPD (Xue et al., 2026) keeps corrections that the student can represent. These methods improve visual privilege, selection, or target construction. Most preserve the realized prefix for alignment. PReViD instead adds an evidence prefix before the unchanged student history while replaying the same response. It does not rewrite the history. A matched placebo prefix filters generic text effects.

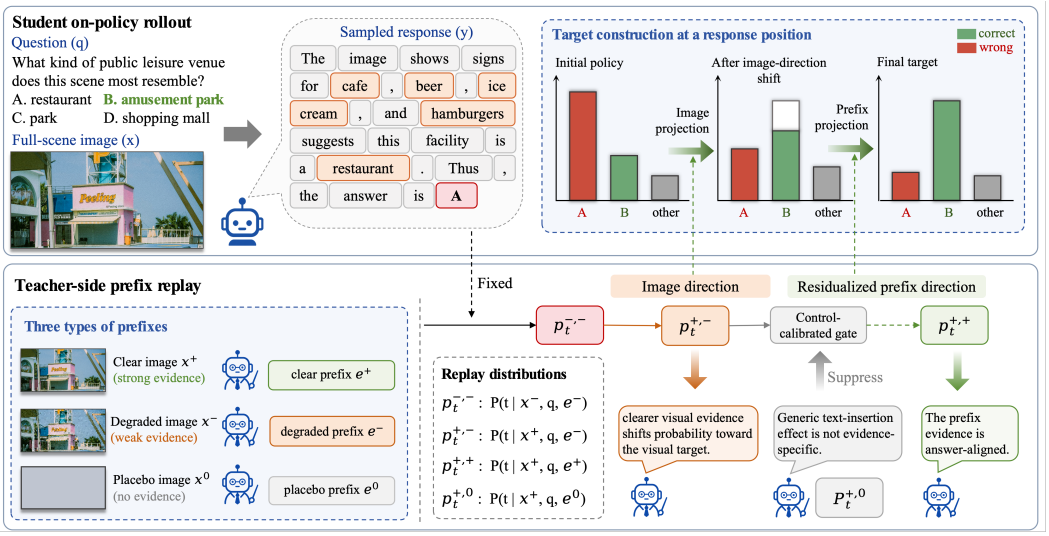


Figure 3: **Overview of PReViD.** The student first samples an on-policy response from the full-scene image. The teacher replays the same response with clear, degraded, and placebo evidence contexts. These comparisons separate the effect of the visual input from the effect of the teacher's evidence context. PReViD uses the selected corrections to construct the training target. Replay is used only during training.

3 METHOD

PReViD adds a training-time intervention to the teacher's evidence context. It keeps the student's on-policy response fixed. As shown in Figure 3, the student first samples one response from the full-scene input. A frozen teacher scores the same response under four selected combinations of image and evidence prefix. The aligned replays define an image direction and an evidence-prefix direction. We remove the image-aligned component from the evidence-prefix direction. A matched placebo direction sets its weight. We project the privileged endpoint correction onto these directions to build a bounded target from the detached student distribution. Auxiliary prefixes and replay are used only during training. Inference uses the original full-scene student and decoding procedure.

3.1 PROBLEM SETUP

We define the student rollout and the evidence prefixes used for teacher replay. Let q be the task prompt and let x be the full-scene student input. Let x^+ and x^- be the clear and degraded evidence inputs. The student samples an on-policy response $y \sim \pi_\theta(\cdot \mid x, q)$. Here, $y = (y_1, \ldots, y_{|y|})$ and $y_{<t}$ is the response history before token y_t. Standard visual OPD evaluates a teacher at every $y_{<t}$. A later target therefore still depends on earlier student errors. PReViD keeps the sampled response fixed and changes the teacher's evidence context used to score it. The tokens in $y_{<t}$ remain unchanged. The evidence prefix is inserted before them.

A frozen observation generator G creates three one-sentence evidence prefixes:

$$e^b = G(x^b, q), \qquad b \in \{+, -, 0\}, \tag{1}$$

Here, x^0 is a uniform RGB placebo image with the same size as the evidence inputs. It is used only to generate the placebo prefix e^0. The student never receives these evidence prefixes.

3.2 INTERVENTIONAL PREFIX REPLAY

We replay the same student response under selected contexts to separate the image and evidence-prefix contributions. Let $p_{\theta,t}$ denote the student's next-token distribution at position t. Let $p_t^{a,b}$ denote the corresponding distribution from the replay teacher π_ϕ. The indices a and b specify the image and

prefix conditions, where $(a, b) \in \{(+,+), (+,-), (-,-), (+,0)\}$. All distributions use the same token support $\mathcal{V}_t$:

$$p_{\theta,t} = \pi_\theta(\cdot \mid x, q, y_{<t}), \qquad p_t^{a,b} = \pi_\phi(\cdot \mid x^a, q, e^b, y_{<t}). \tag{2}$$

The symbols $+$, $-$, and 0 denote clear, degraded, and placebo conditions. The teacher receives each prefix as an unfinished assistant message. It then scores the original response with teacher forcing. Each replay remains aligned with y at every token position. The $(+, 0)$ context keeps the clear image and changes only the prefix to e^0. The four contexts form an ordered path and a placebo comparison. They do not form a complete 2×2 factorial design.

We use lowercase letters for distributions and bold letters for vectors in centered log-probability space. For any distribution p on $\mathcal{V}_t$, define

$$\big[\boldsymbol{\psi}_t(p)\big]_v = \log p(v) - \frac{1}{|\mathcal{V}_t|} \sum_{v' \in \mathcal{V}_t} \log p(v'), \qquad v \in \mathcal{V}_t. \tag{3}$$

Centering removes a common offset and preserves pairwise log-odds.

The ordered path $(x^-, e^-) \to (x^+, e^-) \to (x^+, e^+)$ yields

$$\mathbf{d}_t^x = \boldsymbol{\psi}_t(p_t^{+,-}) - \boldsymbol{\psi}_t(p_t^{-,-}), \qquad \mathbf{d}_t^e = \boldsymbol{\psi}_t(p_t^{+,+}) - \boldsymbol{\psi}_t(p_t^{+,-}). \tag{4}$$

The image direction $\mathbf{d}_t^x$ changes the visual input and holds e^- fixed. The evidence-prefix direction $\mathbf{d}_t^e$ changes the prefix and holds x^+ fixed. Both are controlled model-input contrasts, not estimates of natural direct or indirect effects. The two directions sum to the endpoint contrast $\mathbf{d}_t^x + \mathbf{d}_t^e = \boldsymbol{\psi}_t(p_t^{+,+}) - \boldsymbol{\psi}_t(p_t^{-,-})$. The matched placebo direction is

$$\mathbf{d}_t^0 = \boldsymbol{\psi}_t(p_t^{+,0}) - \boldsymbol{\psi}_t(p_t^{+,-}). \tag{5}$$

It shares the clear image and the degraded-prefix baseline with $\mathbf{d}_t^e$.

3.3 Selective target construction

We use the replay directions to build a selective target distribution for student training. Let $\bar{p}_t$ be a detached copy of $p_{\theta,t}$. The privileged endpoint correction is $\mathbf{c}_t = \boldsymbol{\psi}_t(p_t^{+,+}) - \boldsymbol{\psi}_t(\bar{p}_t)$. For vectors $\mathbf{u}$ and $\mathbf{z}$, define the one-sided ridge projection

$$\Pi_{\mathbf{z}}(\mathbf{u}) = \frac{\mathrm{ReLU}(\langle \mathbf{u}, \mathbf{z} \rangle)}{\|\mathbf{z}\|^2 + \varepsilon} \mathbf{z}, \tag{6}$$

where $\|\cdot\|$ is the Euclidean norm and $\varepsilon > 0$ is a numerical ridge. The ReLU keeps only components aligned with $\mathbf{z}$.

We remove the image-aligned component from the evidence-prefix direction. This avoids counting the image direction twice:

$$\widetilde{\mathbf{d}}_t^e = \mathbf{d}_t^e - \frac{\langle \mathbf{d}_t^e, \mathbf{d}_t^x \rangle}{\|\mathbf{d}_t^x\|^2 + \varepsilon} \mathbf{d}_t^x. \tag{7}$$

The ridge term stabilizes the denominator, so $\widetilde{\mathbf{d}}_t^e$ need not be exactly orthogonal to $\mathbf{d}_t^x$.

The placebo direction is not added to the target. Its raw norm provides a conservative reference for the evidence-prefix direction. Residualizing the placebo direction would reduce this reference and increase the gate. We define the gate as

$$g_t = \sigma\Big(\|\widetilde{\mathbf{d}}_t^e\| - \|\mathbf{d}_t^0\| \Big), \tag{8}$$

where σ is the logistic sigmoid. We use zero margin and unit logistic scale. The gate has no learnable parameters. The one-sided projections retain only components that agree with the privileged endpoint correction. The target shift and target distribution are

$$\begin{aligned} \boldsymbol{\Delta}_t &= \Pi_{\mathbf{d}_t^x}(\mathbf{c}_t) + g_t \Pi_{\widetilde{\mathbf{d}}_t^e}(\mathbf{c}_t), \\ r_t &= \mathrm{softmax}(\log \bar{p}_t + \mathrm{clip}(\boldsymbol{\Delta}_t, -M, M)). \end{aligned} \tag{9}$$

The logarithm and clipping act coordinate-wise. The constant $M > 0$ bounds each log-probability shift. Centering $\log \bar{p}_t$ does not change r_t. Softmax is invariant to a shared additive offset.

3.4 TRAINING OBJECTIVE

We train the student on this target and add a weak anchor to the privileged endpoint. For distributions p and r, define the Jensen–Shannon divergence as $D_{\mathrm{JS}}(p\|r) = \frac{1}{2}D_{\mathrm{KL}}(p\|\mu) + \frac{1}{2}D_{\mathrm{KL}}(r\|\mu)$ with $\mu = (p+r)/2$. The objective is

$$\mathcal{L} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} w_t \left[D_{\mathrm{JS}}(p_{\theta,t}\|r_t) + \lambda D_{\mathrm{JS}}(p_{\theta,t}\|p_t^{+,+})\right], \tag{10}$$

Here, $\mathcal{T}$ is the set of valid response positions. The term w_t is the stored-rollout importance weight. The scalar λ is the anchor weight. The anchor directly regularizes the student toward the privileged endpoint. It is separate from the path decomposition and is not gated. We treat all replay distributions, $\bar{p}_t$, and r_t as constants during optimization. Inference uses only the standard full-scene student.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Backbones and training data. We use Qwen3.5 at the 4B and 9B scales (Qwen Team, 2026). The baseline refers to the original checkpoint before any training. Each trained student is initialized from the corresponding scale-matched baseline. The main comparison uses a scale-matched teacher for each distillation method. We train on the 6,241 visual question and answer pairs from Vision-OPD-6K (Yuan et al., 2026). Each example provides a full-scene student view and a clear evidence crop. The student uses the full-scene view during both training and inference. The clear crop is privileged and is available only during training. We obtain the degraded evidence crop from the clear crop with a fixed transformation. It is therefore a counterfactual evidence view, not a degraded full-scene image. PReViD samples one student response per prompt and replays that response under four teacher contexts. Appendix A.1 lists the exact inputs used by each method.

Baselines. We compare with SFT, GRPO (Shao et al., 2024), and OPSD (Penaloza et al., 2026). We also include VA-OPD (Liu et al., 2026), V-Zero (Sun et al., 2026), Vision-OPD (Yuan et al., 2026), VGS (Yoon et al., 2026), and VAD (Zhang et al., 2026). The fixed-teacher study also includes standard OPD (Agarwal et al., 2024). SFT and GRPO use answer labels without privileged images. GRPO uses a group size of 8. OPSD receives the answer label as textual privileged information. The scale-matched visual OPD methods use the clear crop as privileged visual evidence. We include Gemini 3.5 Flash (Google DeepMind, 2026) and GPT-5.4 (OpenAI, 2026) as reference points. They are not controlled baselines, but we evaluate both using the same benchmark and grading protocol.

Benchmarks and evaluation. We evaluate six datasets from four benchmark suites. V* Bench (Wu & Xie, 2024) tests visual search and detailed grounding in crowded high-resolution images. ZoomBench (Wei et al., 2026) covers fine-grained counting, OCR, object attributes, and object recognition. HR-Bench (Wang et al., 2025) tests fine-grained perception within and across objects at 4K and 8K resolutions. MME-RealWorld (Zhang et al., 2025) evaluates perception and reasoning in high-resolution real-world scenes. We use its English and Chinese splits. We use each benchmark's native answer preprocessing and accuracy metric. We set the decoding temperature to 0 to avoid sampling randomness. ZoomBench requires semantic grading. We use a fixed Doubao-Seed-2.0-Lite (Volcengine, 2026) judge configuration. No evaluation examples are used for training or checkpoint selection. We conduct training and inference on a cluster with 16 NVIDIA H20 GPUs, each with 96 GB of memory. Further implementation details are provided in Appendix A.

4.2 MAIN RESULTS

Table 1 reports the main results. For all methods, the teacher is the same backbone model trained with privileged information. SFT and GRPO do not use privileged information. They improve some benchmarks, but the gains are not consistent across tasks. Visual distillation gives a stronger signal. Methods with privileged information improve over their counterparts without privileged information on several benchmarks. OPSD uses textual privilege, whereas Vision-OPD uses visual privilege. Their relative performance changes with scale. OPSD performs better at 4B, whereas Vision-OPD

Table 1: **Main results on six fine-grained visual benchmarks.** All values are accuracies (%). Avg is the unweighted mean of the six columns. Bold and underline mark the best and second-best results within each Qwen3.5 scale. Closed-source models are reference points and are not included in this ranking. Parentheses in the Ours rows give absolute gains over the corresponding Baseline for individual benchmarks.

Model	V*		Zoom		HR4K		HR8K		MME-EN		MME-CN		Avg
Closed-source models													
Gemini 3.5 Flash	90.58		62.60		87.25		86.63		75.31		72.93		79.22
GPT-5.4	84.82		59.29		85.88		80.00		72.62		70.14		75.46
Qwen3.5-based models: 4B scale													
Baseline	80.63		47.81		83.50		80.00		62.70		63.65		69.72
SFT	81.15		49.59		<u>83.63</u>		78.75		56.33		59.62		68.18
GRPO	79.58		54.79		82.63		78.75		66.56		59.83		70.36
OPSD	82.20		52.07		82.13		<u>80.25</u>		<u>70.65</u>		<u>69.11</u>		72.74
VA-OPD	85.34		53.14		77.38		75.50		67.64		68.90		71.32
V-Zero	84.82		52.43		78.63		75.38		68.18		68.29		71.29
Vision-OPD	86.91		53.25		79.38		77.50		67.16		68.04		72.04
VGS	88.48		<u>57.51</u>		81.63		79.25		67.54		66.81		<u>73.54</u>
VAD	<u>90.05</u>		52.78		77.50		75.25		69.89		67.94		72.24
Ours	**91.10**	(+10.47)	**61.30**	(+13.49)	**84.38**	(+0.88)	**80.63**	(+0.63)	**71.92**	(+9.22)	**70.24**	(+6.59)	**76.60**
Qwen3.5-based models: 9B scale													
Baseline	86.39		52.19		84.25		80.25		70.83		67.50		73.57
SFT	87.43		52.31		84.13		81.88		71.54		68.21		74.25
GRPO	85.86		55.98		84.50		82.50		71.51		68.90		74.88
OPSD	89.01		55.50		83.50		81.38		72.44		68.58		75.07
VA-OPD	90.58		59.88		83.88		82.38		71.82		<u>70.26</u>		76.47
V-Zero	90.05		56.21		83.75		83.88		<u>73.08</u>		69.63		76.10
Vision-OPD	90.58		58.46		84.75		82.63		71.30		68.09		75.97
VGS	89.53		<u>61.18</u>		84.50		83.25		72.23		69.70		76.73
VAD	<u>92.67</u>		57.99		<u>85.38</u>		**85.50**		72.28		69.38		<u>77.20</u>
Ours	**93.72**	(+7.33)	**64.50**	(+12.31)	**86.13**	(+1.88)	<u>84.88</u>	(+4.63)	**74.43**	(+3.60)	**70.39**	(+2.89)	**79.01**

performs better at 9B. Greater visual capacity may help the larger model use the evidence crop more effectively. It suggests that neither source of privileged information is uniformly stronger. PReViD therefore models both visual and contextual corrections instead of relying on privileged vision alone.

Existing visual OPD methods mainly refine the visual target. Later tokens are still conditioned on the student's previous tokens. With a fixed history, early mistakes can leave a corrupted language context. A better visual target alone cannot fully correct this context. Prefix replay mitigates this issue by placing an evidence prefix before the unchanged student history. The prefix provides an alternative visual statement that can reduce the effect of earlier mistakes. Among the scale-matched Qwen3.5-based methods, PReViD gives the best average performance at both model scales and leads on most benchmarks. This suggests that the evidence-prefix effect is not tied to a single task.

4.3 OPD WITHOUT ADDITIONAL PRIVILEGED FIELDS

Table 1 studies scale-matched distillation with privileged training views. Such data often require manual annotation or generation by strong models, followed by accuracy checks. Motivated by this limitation, we test whether PReViD can transfer knowledge from a stronger fixed teacher without providing these fields again during 4B student distillation. We assume access to a strong teacher and use the best 9B checkpoint from Table 1, namely Ours. This teacher was trained with privileged views in the main experiment. All trained students are initialized from Qwen3.5-4B. We also compare OPD, VA-OPD, and VGS. During 4B student training, these methods do not use answer labels or privileged fields. Appendix Table 5 lists their training inputs.

Table 2 reports results under this fixed-teacher setting. All methods use the same student initialization and teacher checkpoint. They differ in how they extract supervision from the teacher distribution. A stronger teacher alone does not ensure effective transfer. The contrastive baselines improve most benchmarks, but their gains are limited. Vanilla OPD transfers knowledge more consistently. It still

Table 2: **OPD without additional privileged fields during 4B student distillation.** All trained models are Qwen3.5-4B students. The second row reports the fixed Qwen3.5-9B PReViD teacher and is excluded from the student ranking. Bold and underline mark the best and second-best student results. Parentheses in the Ours row give absolute gains over the 4B Baseline for individual benchmarks. Avg is the unweighted mean.

Method or role	V*	Zoom	HR4K	HR8K	MME-EN	MME-CN	Avg
Baseline	80.63	47.81	83.50	<u>80.00</u>	62.70	63.65	69.72
9B teacher (Ours)	93.72	64.50	86.13	84.88	74.43	70.39	79.01
OPD	87.43	<u>53.02</u>	<u>83.88</u>	79.50	<u>69.71</u>	<u>68.43</u>	<u>73.66</u>
VA-OPD	<u>87.96</u>	49.82	82.38	77.13	67.86	67.89	72.17
VGS	86.91	48.52	79.88	75.63	66.53	66.01	70.58
Ours	**88.48** (+7.85)	**59.64** (+11.83)	**84.88** (+1.38)	**82.13** (+2.13)	**73.64** (+10.94)	**70.24** (+6.59)	**76.50**

conditions on the uncorrected student history. Ours improves more clearly over vanilla OPD and further narrows the gap to the teacher. Without additional privileged fields at the student stage, the distillation gains of the other methods are less pronounced. Our method achieves the best result among all student models. This suggests that PReViD is not limited to privileged-view distillation. It can also distill a fixed teacher without reusing answer labels or privileged fields at the student stage. Similar to OPD, PReViD preserves the student's on-policy trajectory. PReViD adds teacher-generated evidence before the unchanged history. This strengthens later-token supervision.

4.4 Method ablations

As in the main experiments, all ablations use Qwen3.5-4B students and a same-architecture teacher trained with privileged visual information. We start from Vision-OPD. Vision-OPD follows a training form similar to OPSD. It replaces textual privileged information with visual privileged information and supervises the full sequence. Table 3 then adds the components of PReViD step by step.

Visual projection keeps only the teacher correction aligned with the image direction. This changes the supervision from a global token-level constraint to a direction that better reflects the teacher's visual focus. The remaining variants add the evidence-prefix path on top of visual projection. They differ only in how the prefix path is selected. The no-gate variant applies all prefix shifts. The random-gate control replaces the deterministic gate with random selection. The shuffled-prefix control uses prefix information from another training example. Full PReViD uses the gate to weight the matched evidence prefix.

Table 3: **Ablation study of individual components.** Rows start from Vision-OPD and add visual projection. The four evidence-prefix variants below the rule share the same visual projection base. They differ only in how the evidence-prefix path is selected. Bold and underline mark the best and second-best results. Parentheses in the Ours row show absolute gains over the Baseline on each benchmark. Avg denotes the unweighted mean.

Training variant	V*	Zoom	HR4K	HR8K	MME-EN	MME-CN	Avg
Baseline	80.63	47.81	83.50	<u>80.00</u>	62.70	63.65	69.72
Vision-OPD	86.91	53.25	79.38	77.50	67.16	68.04	72.04
+ visual projection	87.43	54.91	83.38	77.13	66.84	68.46	73.03
Evidence-prefix path added to visual projection							
all prefix shifts (no gate)	<u>89.01</u>	<u>58.82</u>	<u>84.00</u>	79.25	<u>68.06</u>	<u>69.65</u>	<u>74.80</u>
random gate	87.96	57.40	83.63	78.75	67.38	66.94	73.68
shuffled prefix	84.29	55.86	82.25	78.25	64.76	65.13	71.76
Ours	**91.10** (+10.47)	**61.30** (+13.49)	**84.38** (+0.88)	**80.63** (+0.63)	**71.92** (+9.22)	**70.24** (+6.59)	**76.60**

The ablation results are shown in Table 3. This isolates the contribution of each design choice. Compared with Vision-OPD, which uses privileged visual information, visual projection improves average performance. This suggests that projecting the teacher correction onto the image direction provides effective supervision. Adding the evidence-prefix path brings a larger gain. Full PReViD

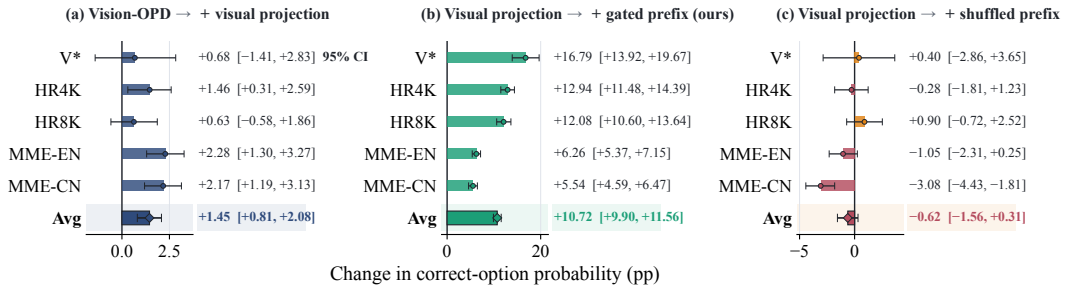


Figure 4: **The gated evidence-prefix path gives the largest and most stable increase in correct-option probability.** Each panel reports the mean paired change in correct-option probability across five benchmarks, measured in percentage points. (a) Visual projection over Vision-OPD. (b) Gated prefix over visual projection. (c) Shuffled prefix over visual projection. Error bars show 95% confidence intervals. Avg denotes the unweighted mean over the five benchmarks.

achieves the best result on every benchmark. PReViD outperforms the no-gate and random-gate variants. This shows that contextual correction should not be applied uniformly. It should be used selectively based on the current example. PReViD also outperforms the shuffled-prefix variant. This shows that the correction signal must come from evidence grounded in the current example. An arbitrary prefix does not provide a reliable gain. These results support two key motivations of our method: recovering the missing contextual correction and transferring it only when it is relevant.

Table 3 mainly reports answer accuracy for each variant on each benchmark. We provide a complementary probability-based analysis in Figure 4. This analysis examines how the image path and the evidence-prefix path change the probability assigned to the correct option, including changes that do not alter the decoded answer. ZoomBench is excluded because its semantic grading does not define a fixed option set. Visual projection produces a positive shift in correct-option probability, but the gain is small. The effect also varies across benchmarks. The confidence intervals quantify uncertainty in the mean paired change, rather than variation across individual examples. These intervals show that visual projection does not consistently increase the model's probability of producing the correct answer. In contrast, the gated evidence-prefix path produces a clear positive shift on every benchmark. All confidence intervals lie entirely above zero, indicating a more stable increase in correct-option probability. A shuffled prefix does not yield a reliable improvement and produces unstable changes. This further shows that adding a prefix alone does not reliably improve model predictions.

Therefore, PReViD not only improves benchmark accuracy but also increases the probability assigned to the correct option more consistently. This result is consistent with the accuracy gains in Table 3. It shows that the improvement comes from the selective transfer of relevant evidence, rather than from the presence of extra text alone.

5 Conclusion

Fixed-history visual OPD can correct an early token, but it does not account for how that correction should affect later predictions. PReViD recovers this missing supervision signal by replaying the same student response under controlled image and evidence contexts. It separates image and evidence-prefix corrections, uses a matched placebo prefix to filter generic text effects, and selectively transfers context changes that support the teacher's visual correction. Across six fine-grained visual benchmarks, PReViD achieves the best average performance among the Qwen3.5-based methods at both 4B and 9B scales. It also performs best among 4B students trained by a fixed 9B teacher without additional privileged fields during distillation. Component ablations support the benefits of visual projection, matched evidence prefixes, and selective gating. Probability-based analysis shows that the gated evidence-prefix path increases the probability of the correct option more consistently.

Limitations. PReViD mitigates the effect of incorrect evidence in a fixed student history, but it does not replace that history with a corrected rollout. Future work could jointly revise the response and its context. These replay operations are used only during training, and although they can be parallelized, they may still introduce a small amount of additional training time.

AI USE STATEMENT

We used generative AI tools only for grammar checking, minor text editing, and code completion. We did not use these tools for any other tasks that require disclosure. We reviewed all AI-assisted work and take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

We release our code through an anonymous repository. It includes detailed configurations for data processing, model training, inference, and evaluation. We use the same training data and benchmarks as related work. We do not apply generative modifications to these datasets, which keeps the data setting consistent across methods. We fix the random seed and set the decoding temperature to 0 to reduce randomness.

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*, 2024.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. Opd-v: Visual on-policy self-distillation with modality balance, 2026. URL https://arxiv.org/abs/2608.05131.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Lin Chen, Xiaoke Zhao, Kun Ding, Weiwei Feng, Changtao Miao, Zili Wang, Wenxuan Guo, Ying Wang, Kaiyuan Zheng, Bo Zhang, Zhe Li, and Shiming Xiang. Beyond next-token alignment: Distilling multimodal large language models via token interactions, 2026. URL https://arxiv.org/abs/2602.09483.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi WANG, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, and Libo Qin. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, June 2025.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9062–9072, June 2025.

Google DeepMind. Gemini 3.5 flash: Model card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026.

Yuxuan Jiang and Francis Ferraro. Bridging reasoning trajectories in on-policy distillation via near-future guidance, 2026. URL https://arxiv.org/abs/2606.00305.

Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Vineeth N. Balasubramanian, and Tanuja Ganu. Chain-of-thought degrades visual spatial reasoning capabilities of multimodal LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 862–876, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-391-3. doi: 10.18653/v1/2026.acl-short.71.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. Compodistill: Attention distillation for compositional reasoning in multimodal LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation, 2016. URL https://arxiv.org/abs/1606.07947.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

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. In *Forty-third International Conference on Machine Learning*, 2026a.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-supervised visual on-policy distillation, 2026b. URL https://arxiv.org/abs/2608.14144.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation, 2026. URL https://arxiv.org/abs/2607.21556.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models, 2026. URL https://arxiv.org/abs/2605.21924.

OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. In *Forty-third International Conference on Machine Learning*, 2026.

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

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

Fangxun Shu, Yue Liao, Lei Zhang, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Long Chan, TaoZhong, Zhelun Yu, Wanggui He, Siming Fu, Haoyuan Li, Si Liu, Hongsheng Li, and Hao Jiang. LLaVA-mod: Making LLaVA tiny via moe-knowledge distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haoxiang Sun, Zhihang Yi, Langxuan Deng, Yuhao Zhou, Peiqi Jia, Jian Zhao, Li Yuan, Jiancheng Lv, and Tao Wang. V-zero: Answer-label-free on-policy distillation with contrastive evidence gating for fine-grained visual reasoning, 2026. URL https://arxiv.org/abs/2606.25319.

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

Daniil Tiapkin, Daniele Calandriello, Johan Ferret, Sarah Perrin, Nino Vieillard, Alexandre Rame, and Mathieu Blondel. On teacher hacking in language model distillation. In *Forty-second International Conference on Machine Learning*, 2025.

Volcengine. Model list: Doubao-seed-2.0-lite. https://www.volcengine.com/docs/82379/1799865, 2026.

Shuai Wang, Daoan Zhang, Zhe Tang, Hao Cheng, and Jiaheng Wei. Self-boosting vision-language models with noisy student on-policy self-distillation, 2026. URL https://arxiv.org/abs/2607.23125.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: a training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 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. In *Forty-third International Conference on Machine Learning*, 2026.

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

Feng Xiong, Leyan Xue, and Hongyu Lin. Correcting what you cannot see: Credit assignment for perception distillation in multimodal reasoners, 2026. URL https://arxiv.org/abs/2607.28336.

Wenda Xu, Rujun Han, Zifeng Wang, Long 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. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.14084.

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

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark A. Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. In *Forty-third International Conference on Machine Learning*, 2026.

Seonghoon Yu, Dongjun Nam, Byung-Kwan Lee, and Jeany Son. Hide to see: Reasoning-prefix masking for visual-anchored thinking in vlm distillation, 2026. URL https://arxiv.org/abs/2605.11651.

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.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. Vad: Attributing visual evidence for target reconstruction in multimodal on-policy distillation, 2026. URL https://arxiv.org/abs/2607.28590.

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*, 2025.

Wenhong Zhu, Ruobing Xie, Rui Wang, and Pengfei Liu. Hybrid policy distillation for LLMs. In *Forty-third International Conference on Machine Learning*, 2026a.

Zhihao Zhu, Jiafeng Liang, Shixin Jiang, Jinlan Fu, Ming Liu, Guanglu Sun, See-Kiong Ng, and Bing Qin. Analyzing reasoning consistency in large multimodal models under cross-modal conflicts, 2026b. URL https://arxiv.org/abs/2601.04073.

Appendix

A Implementation and training details

A.1 Training inputs and supervision

Vision-OPD-6K directly provides the question, full image, evidence crop, and answer label. We generate the remaining visual inputs by script. The degraded image is obtained by applying a fixed blur to the evidence crop. The non-evidence crop is randomly sampled from a region outside the evidence box. The placebo image is a uniform RGB image with the same dimensions as the crop. These three derived inputs require no additional annotation and can be generated quickly. Table 4 lists the inputs and supervision used by each training method in the main results.

Table 4: **Training inputs and supervision for the main results in Table 1.** All students receive only the question and full image at inference.

Method	Training inputs and supervision
Without privileged visual evidence	
SFT	Question, full image, answer label
GRPO	Question, full image, answer label for reward
OPSD	Question, full image, textual privileged information (answer label)
With privileged visual evidence	
VA-OPD	Question, full image, evidence crop, degraded image
V-Zero	Question, full image, evidence crop, non-evidence crop
Vision-OPD	Question, full image, evidence crop
VGS	Question, full image, evidence crop, degraded image
VAD	Question, full image, evidence crop, degraded image
PReViD	Question, full image, evidence crop, degraded image, placebo image

The experiment in Section 4.3 uses no answer labels or privileged visual evidence. The full image serves as the clear teacher view. We apply the same fixed transformation to obtain the degraded image. The placebo image is a uniform RGB image with the same dimensions as the full image.

Table 5: **Training inputs for the fixed-teacher experiment in Section 4.3.** No method uses answer labels or privileged visual evidence.

Method	Training inputs
OPD	Question, full image
VA-OPD	Question, full image, degraded image
VGS	Question, full image, degraded image
PReViD	Question, full image, degraded image, placebo image

A.2 Evidence-prefix generation and interventional replay

For each example, the generator receives the question and one of three visual conditions: the evidence crop, the degraded image, or the placebo image. The placebo image is x^0 from Section 3. For scale-matched experiments, G is a frozen copy of the corresponding Baseline and also serves as the replay teacher. In the fixed-teacher experiment, the frozen 9B PReViD checkpoint serves both roles. The generator does not receive the answer label or the sampled student response. We use the following prompt for all three conditions:

Evidence-prefix generation prompt

```
{question}
Before giving the requested answer, write exactly one short analysis
sentence that explicitly states the concrete visual fact you observed.
Stop after that sentence; do not emit a separate final-answer field
and do not repeat the choices.
```

This procedure produces e^+, e^-, and e^0 in Eq. 1. Generation is greedy and each prefix contains at most 64 tokens. For each training batch, the three prefix-generation calls run in parallel with the student rollout. The generated prefixes are then reused by the teacher replays. The prompt forbids a final-answer field and choice repetition. A visual fact can still be answer-relevant. We therefore treat the clear prefix as generated teacher supervision, not as an answer-independent signal. The student never receives them at inference. During replay, the teacher receives a prefix as an unfinished assistant message. It then scores the same sampled response with teacher forcing. The four contexts therefore remain aligned at every response position.

A.3 Shared token support and numerical stability

The student and replay distributions share the same token support. At each response position, the current student selects $K = 64$ vocabulary tokens. We evaluate every replay teacher on the same token IDs. One extra coordinate stores the probability mass outside this set. We compute this mass with a stable log-sum-exp residual. We then center the log probabilities on the shared support $\mathcal{V}_t$.

A.4 Choice of training divergence

PReViD uses JSD for the two matching terms in Eq. 10. We compare it with forward and reverse KL while keeping all other settings fixed. For either target $z \in \{r_t, p_t^{+,+}\}$, forward KL uses $D_{\mathrm{KL}}(z \| p_{\theta,t})$. Reverse KL uses $D_{\mathrm{KL}}(p_{\theta,t} \| z)$.

The three divergences give similar results. Their average accuracies span 0.49 points at 4B and 0.73 points at 9B. The best divergence varies across individual benchmarks. JSD gives the highest average at both scales. It exceeds reverse KL by 0.06 points at 4B and 0.18 points at 9B. We therefore use JSD as the default objective.

Table 6: **Effect of the training divergence.** All values are accuracies (%). Bold and underline mark the best and second-best divergence results within each model scale. Avg is the unweighted mean of the six benchmarks.

Training objective	V*	Zoom	HR4K	HR8K	MME-EN	MME-CN	Avg
Qwen3.5-based models: 4B scale							
Baseline	80.63	47.81	83.50	80.00	62.70	63.65	69.72
Forward KL	90.05	61.07	84.13	**80.75**	71.51	69.16	76.11
Reverse KL	**91.62**	60.95	**84.50**	80.63	71.71	69.85	76.54
JSD (Ours)	91.10	**61.30**	84.38	80.63	**71.92**	**70.24**	**76.60**
Qwen3.5-based models: 9B scale							
Baseline	86.39	52.19	84.25	80.25	70.83	67.50	73.57
Forward KL	92.67	63.79	85.25	84.13	74.28	69.55	78.28
Reverse KL	**93.72**	64.26	85.88	**85.00**	74.11	70.02	78.83
JSD (Ours)	**93.72**	**64.50**	**86.13**	84.88	**74.43**	**70.39**	**79.01**

A.5 Training cost

We compare all methods with Qwen3.5-4B for one epoch on Vision-OPD-6K. GPU hours equal wall-clock training time multiplied by the number of GPUs. They include student rollout, concurrent prefix generation, four teacher replays, and optimization. Prefix generation runs in parallel with the student rollout. The student rollout dominates this stage, so prefix generation adds only a small wall-clock increase. We also pipeline the teacher replays to improve GPU utilization. Vision-OPD has the lowest cost among the OPD methods. Its compact evidence crop may reduce teacher-side visual encoding time. PReViD costs $1.07\times$ Vision-OPD, close to comparable OPD methods. GRPO costs $4.15\times$ Vision-OPD. The replay overhead is therefore modest in this setting.

Table 7: **Training cost under the standard Vision-OPD-6K setting.** All methods use Qwen3.5-4B and train for one epoch on the same data. GRPO uses a group size of 8. Relative cost is normalized to Vision-OPD.

Method	GPU hours	Relative cost
SFT	5.59	$0.21\times$
GRPO	108.11	$4.15\times$
OPSD	27.54	$1.06\times$
VA-OPD	26.75	$1.03\times$
V-Zero	32.41	$1.25\times$
Vision-OPD	26.03	$1.00\times$
VGS	29.62	$1.14\times$
VAD	26.83	$1.03\times$
PReViD	27.77	$1.07\times$

B Full-corpus ordered-path analysis

Figure 2 measures the ordered image and evidence-prefix paths through controlled replay. Table 3 tests their effect on downstream accuracy. Here, we define the scalar score used in this analysis and report its uncertainty.

We evaluate the ordered intervention in Eq. 4 with Qwen3.5-4B on all 6,241 Vision-OPD-6K examples. At the scored answer position t, let $v^\star \in \mathcal{V}_t$ be the correct candidate. Let $\mathcal{F}_t \subset \mathcal{V}_t$ contain the incorrect candidates. Using the replay distributions $p_t^{a,b}$ from Eq. 2, we define the target-versus-foil log-odds as

$$\ell_t^{a,b} = \log p_t^{a,b}(v^\star) - \log \sum_{v \in \mathcal{F}_t} p_t^{a,b}(v). \tag{11}$$

The common softmax normalizer cancels. Thus, this score also equals the correct-candidate logit minus the log-sum-exp over foil logits. The scalar image and evidence-prefix path components are

$$\delta_t^x = \ell_t^{+,-} - \ell_t^{-,-},$$
$$\delta_t^e = \ell_t^{+,+} - \ell_t^{+,-}.$$

They correspond to the vector directions $\mathbf{d}_t^x$ and $\mathbf{d}_t^e$ in Eq. 4. Their sum is the endpoint contrast $\delta_t^x + \delta_t^e = \ell_t^{+,+} - \ell_t^{-,-}$.

We also report a visually sensitive subset. It contains examples whose highest-probability candidate is correct under $p_t^{+,+}$ and incorrect under $p_t^{-,-}$. Selection therefore depends on the joint image-prefix endpoints, not on an image-only change. "Visually sensitive" is shorthand for this behaviorally selected subset. Its results describe the mechanism within that subset and are not a population-level effect estimate.

Table 8: **Ordered-path log-odds decomposition on Vision-OPD-6K.** Values are means with 95% bootstrap confidence intervals over examples.

Stratum	n	Endpoint contrast	Image path	Evidence-prefix path
All examples	6,241	3.44 [3.27, 3.62]	0.82 [0.78, 0.86]	2.62 [2.47, 2.77]
Visually sensitive	2,077	11.69 [11.52, 11.85]	2.17 [2.10, 2.23]	9.52 [9.36, 9.68]

Across all examples, the evidence-prefix path accounts for 2.62 of the 3.44 mean endpoint contrast, or 76%. This share rises to 81% in the visually sensitive subset. The larger component in this subset shows that path strength varies across examples. This result motivates selective transfer. It does not by itself establish the downstream benefit of the gate.

C Controlled example of interventional prefix replay

We use a red triangle x^{R} and a matched blue triangle x^{B}. This diagnostic is separate from Table 8. The superscripts R and B are local color labels. The blue input is a foil, not the degraded evidence view x^- used in the full-corpus analysis.

Greedy observation generation gives

e^{R}: "The large shape in the image is a solid red triangle."
e^{B}: "The large shape in the image is a blue triangle."

For each image-prefix pair, we append "Based on this visual observation, the one-word answer is". One teacher pass gives the next-token distribution at position t. We retain the context notation $p_t^{a,b}$ from Eq. 2, with local indices $a, b \in \{\mathrm{R}, \mathrm{B}\}$. The target-versus-foil score in Eq. 11 then becomes

$$\ell_t^{a,b} = \log p_t^{a,b}(\texttt{red}) - \log p_t^{a,b}(\texttt{blue}). \tag{12}$$

This is the two-candidate case of Eq. 11.

Panel (a) of Figure 5 reports all four controlled contexts. The diagonal pairs have matching images and prefixes. The off-diagonal pairs place the image and prefix in conflict.

The ordered path uses three contexts: $(x^{\mathrm{B}}, e^{\mathrm{B}}) \to (x^{\mathrm{R}}, e^{\mathrm{B}}) \to (x^{\mathrm{R}}, e^{\mathrm{R}})$. The first step changes only the image. The second step changes only the evidence prefix:

$$\delta_t^x = \ell_t^{\mathrm{R,B}} - \ell_t^{\mathrm{B,B}} = 2.500,$$
$$\delta_t^e = \ell_t^{\mathrm{R,R}} - \ell_t^{\mathrm{R,B}} = 14.125,$$
$$\delta_t^{\mathrm{end}} = \ell_t^{\mathrm{R,R}} - \ell_t^{\mathrm{B,B}} = 16.625. \tag{13}$$

Panel (b) shows this ordered decomposition. The evidence-prefix path accounts for 14.125 of the 16.625 endpoint contrast.

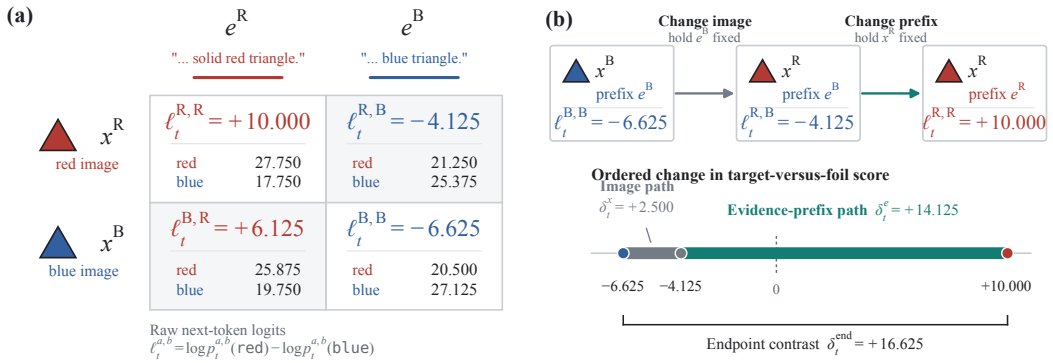


Figure 5: **Controlled contexts and their ordered decomposition.** All logits and scores are from Qwen3.5-4B with BF16 inference. (a) Rows change the image. Columns change the evidence prefix. Each cell reports the raw next-token logits and the score in Eq. 12. (b) The path changes the image first and the prefix second. Segment lengths show δ_t^x, δ_t^e, and the endpoint contrast.

The fourth pair, $(x^{\mathrm{B}}, e^{\mathrm{R}})$, is outside this ordered decomposition. PReViD uses the matched placebo (x^{+}, e^{0}) as its fourth replay. This controlled comparison does not treat the prefix as a natural causal mediator.