PA-OPSD: ON-POLICY SELF-DISTILLATION JOINTLY GUIDED BY CRITICAL GENERATION POSITIONS AND VISUAL ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) has emerged as a promising paradigm for enhancing multimodal large language models (MLLMs). However, existing methods typically apply indiscriminate distillation across all decoding steps and align only output token distributions. This often leads to two critical pitfalls: over-penalizing non-visual linguistic tokens and encouraging superficial output mimicry without faithful visual grounding. In this work, we propose PA-OPSD (Privileged-Anchor On-Policy Self-Distillation), a framework that unifies generation-position selection and visual evidence alignment via task-relevant visual anchors. Specifically, an exponential moving average (EMA) teacher extracts privileged anchors by aggregating cross-modal attention conditioned on the query, ground-truth answer, and the student's early rollout. These frozen anchors drive two synergistic supervision pathways along student-generated trajectories: (1) Anchor-guided visual contrast, which evaluates teacher divergence under localized visual masking to filter visually sensitive positions while sharpening target distributions via log-probability contrast; and (2) Spatial visual attention distillation, which directly aligns the student's internal visual attention with the anchor-guided teacher exclusively at these critical positions. Privileged answers are utilized solely for anchor extraction during training, leaving the inference pipeline and architecture entirely untouched. Extensive experiments across three base model families and eight diverse benchmarks demonstrate that PA-OPSD consistently outperforms existing OPSD methods, substantially mitigating hallucination and improving fine-grained multimodal reasoning. Code is available in supplementary material.

1 INTRODUCTION

On-policy self-distillation (OPSD) has emerged as an efficient post-training paradigm for advancing the perception and reasoning capabilities of multimodal large language models (MLLMs). Unlike conventional offline distillation that queries static teacher trajectories, on-policy distillation evaluates the teacher model directly on response prefixes rolled out by the student. This dynamic alignment enforces supervision on states the student actually visits, substantially alleviating exposure bias and train-generation distribution shifts (Agarwal et al., 2023). Furthermore, OPSD strengthens the self-teacher with training-time privileged information—such as reference solutions or extended rationale—providing denser, token-level supervisory signals derived from the model's own latent capacity (Zhao et al., 2026a).

To construct an informative discrepancy between student and teacher conditions, recent multimodal self-distillation approaches leverage localized visual perturbations, such as cropping, resolution degradation, or multi-view contrasts (Yuan et al., 2026; Wang et al., 2026b; Liang et al., 2026). While these strategies enrich teacher signals via input-space interventions, a fundamental question remains largely unanswered: *does an augmented teacher advantage naturally ensure that the student absorbs visual evidence effectively?* In practice, constructing an informative teacher is only half the battle. Two deeply intertwined challenges govern the efficacy of multimodal distillation: **where along the generation trajectory to inject supervision**, and **how to ensure the student faithfully exploits the underlying visual evidence at those steps** (Figure 1).

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.

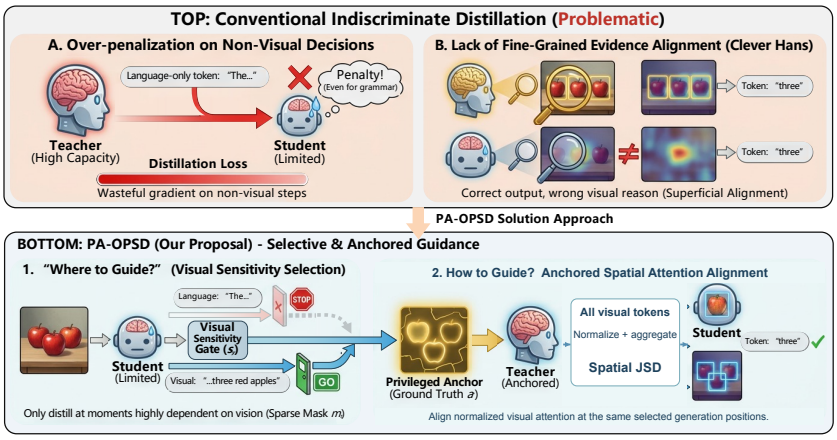

Figure 1: Motivation for PA-OPSD. Uniform distillation may over-penalize nonvisual decisions and align outputs without aligning visual evidence (top). PA-OPSD focuses supervision on visually sensitive positions and aligns spatial attention using privileged anchors (bottom).

First, conventional OPSD frameworks enforce indiscriminate, uniform distillation across all response tokens (Zhao et al., 2026a; Liang et al., 2026), oblivious to the heterogeneous visual dependence of individual decoding steps. Throughout generation, non-visual tokens such as syntactic connectors, formatting markers, and boilerplate reasoning steps are predominantly dictated by linguistic priors. Conversely, critical disambiguation decisions—such as numerical quantities, spatial topologies, and geometric measurements—hinge strictly on fine-grained visual regions. Penalizing the student uniformly on language-dominated decisions not only wastes optimization budget on stylistic and grammatical trivialities, but also dilutes the supervisory gradients at visually sensitive bottlenecks. While recent attempts dynamically reweight supervision via visual contrasts or confidence metrics (Liu et al., 2026; Qian et al., 2026; Aniri et al., 2026), there lacks a principled mechanism that grounds token selection in an explicit, task-relevant visual evidence base.

Second, output distribution matching alone provides no structural guarantee that the student internalizes the teacher's evidence-seeking process. The mapping from internal multimodal attention representations to final vocabulary distributions is fundamentally non-injective; distinct internal attention configurations can produce identical output logits. Consequently, a student may learn superficial shortcuts—arriving at the correct next-token prediction via language context alone while completely ignoring or misinterpreting the visual evidence (the classic "Clever Hans" effect). As benchmarked in visual reasoning suites (Lu et al., 2023; Zhang et al., 2024), true perceptual competence demands not merely predicting the answer, but anchoring reasoning in the correct visual premises. Purely behavioral logit distillation fails to penalize unfaithful reasoning shortcuts, leaving internal cross-modal alignment uncalibrated.

To resolve this dual dilemma, we propose **PA-OPSD** (**P**rivileged-**A**nchor **O**n-**P**olicy **S**elf-**D**istillation), a framework that unifies generation-position selection and internal attention supervision through shared privileged visual anchors. As conceptualized in Figure 1, the core insight of PA-OPSD is to let the privileged teacher ground the task's visual essence into a set of spatial anchors, which simultaneously govern *where to distill* and *what visual evidence to attend to*. Specifically, an exponential moving average (EMA) teacher extracts complementary visual anchors by aggregating cross-modal text-to-visual attention across layers and heads, adaptively conditioned on the query prompt, the student's early rollout prefix, and the training-only ground-truth answer. Using these frozen anchors as a shared coordinate system, PA-OPSD establishes two synergistic supervision pathways along student-generated rollouts:

1. **Visual-Sensitivity-Driven Selective Output Distillation:** By comparing teacher predictions on the original image against a control image where anchor-selected Top-K visual regions are counterfactually masked, we quantify the visual reliance of each generation step via Jensen–Shannon divergence (JSD). This isolates visually sensitive positions (forming a sparse mask m_t) and shapes contrastively enhanced target distributions to sharpen prediction targets.

2. **Spatial Attention Distillation at Critical Positions:** Sharing the identical position mask m_t, we directly align the student's normalized spatial visual attention with that of the anchor-guided teacher via a unified spatial JSD loss. This explicitly forces the student to look at the correct visual regions during critical generation steps, preventing shortcut learning.

Importantly, ground-truth answers are strictly confined to anchor extraction on the teacher side. The teacher receives no privileged text during predictive passes, and inference preserves the student model entirely unchanged in both architecture and deployment efficiency. We evaluate PA-OPSD across three diverse MLLM backbones (Qwen3-VL-8B, Qwen3.5-4B, and Qwen3.5-9B) on eight challenging benchmarks covering multidisciplinary reasoning, fine-grained perception, and hallucination diagnosis. PA-OPSD consistently establishes state-of-the-art self-distillation performance.

2 Related Work

2.1 Multimodal On-Policy Distillation

On-policy distillation (OPD) supervises student-generated trajectories to mitigate distribution shift between training and inference (Agarwal et al., 2023). H-OPD and OPOD transfer complementary capabilities from heterogeneous teachers to a unified student. They use token-level teacher fusion and modality-based routing, respectively (Yin et al., 2026; Zhao et al., 2026b). To improve visual evidence use, VA-OPD and Med-OPD use visual contrasts to reallocate supervision at the token and trajectory levels (Liu et al., 2026; Qian et al., 2026). ADOPD calibrates trajectory weights with matched and mismatched reference images (He et al., 2026). Beyond supervision weighting, VAD and OPD-Aha extract corrective signals from teacher predictions under counterfactual visual conditions and reconstruct the targets (Zhang et al., 2026a; Qiu et al., 2026). FP-OPD uses visual perturbations to estimate locally expressible directions of the student, then projects teacher corrections onto this space (Xue et al., 2026). OPLD jointly aligns output distributions and latent reasoning trajectories, extending supervision to internal reasoning processes (Zhu et al., 2026b).

2.2 Multimodal On-Policy Self-Distillation

On-policy self-distillation (OPSD) derives supervision from condition-specific branches, frozen copies, or EMA copies of the same model. Reference solutions can provide training-time privileges (Zhao et al., 2026a). Visual-OPSD transfers reasoning from visual thought trajectories to a text-only path (Li et al., 2026a). Vision-OPD, Clue-OPSD, and Video-OPSD focus on local regions or key temporal segments (Yuan et al., 2026; Wang et al., 2026a;c). NOPD, S^2VOPD, and RP-OPSD exploit visual perturbations or resolution gaps (Wang et al., 2026b; Li et al., 2026b; Zhu et al., 2026a). ViCuR instead recovers task-relevant visual cues without answer privileges, supporting both OPD and OPSD (Tian et al., 2026). VCSD and CVPD use visual contrasts to refine targets and identify perceptual blind spots, respectively (Liang et al., 2026; Venkatraman et al., 2026). Video-OPSD, OPD-V, and GUI-SD weight or select supervision using visual dependence, modality balance, and teacher confidence, respectively (Wang et al., 2026c; Aniri et al., 2026; Zhang et al., 2026b). BAS-OPD selects entire rollouts under a query budget (Chen et al., 2026). PA-OPSD extracts regional anchors from the question, training-only answer, and early response. Predictions on the original image and a control image with the selected Top-K regions masked provide both JSD-based position selection and VCSD-style contrastive target shaping. Drawing on ADAPT's attention guidance (Yao et al., 2026), we also align spatial attention at the selected positions using frozen anchors. Unlike OPLD's latent-trajectory alignment, PA-OPSD jointly aligns critical response positions and visual regions. Answers enter only anchor extraction; the teacher never receives answer text during prediction. Inference retains only the student.

3 Method

3.1 Problem Formulation and Overview of PA-OPSD

Consider a multimodal supervised dataset $\mathcal{D} = \{(I_i, x_i, a_i^\star)\}_{i=1}^N$. Here, I_i, x_i, and $a_i^\star$ denote the image, question, and ground-truth answer available only during training, respectively. The dataset contains N examples. We omit the example index below and write a single example as $(I, x, a^\star)$.

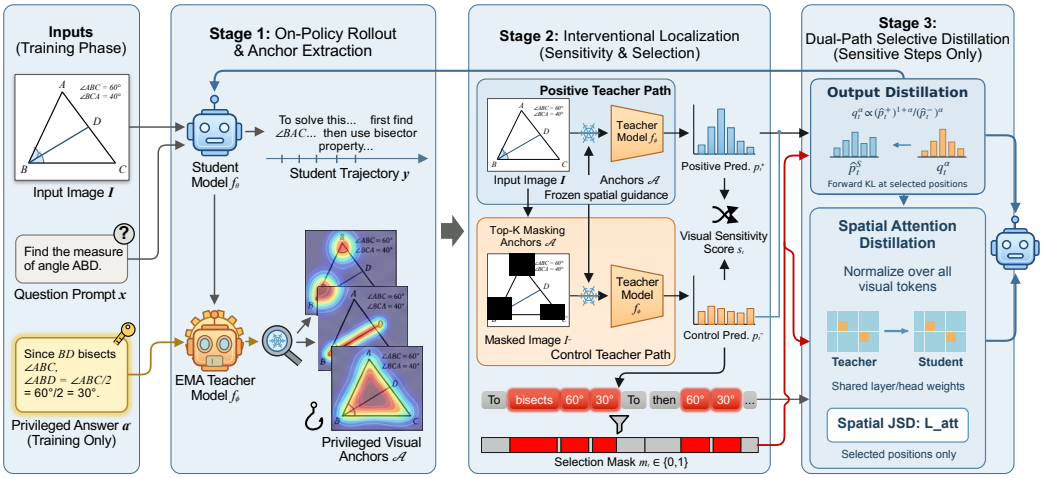

Figure 2: Overview of PA-OPSD. The EMA teacher extracts privileged anchors from the question, training-only answer, and early response. Predictions on original and region-masked images select critical positions and form contrastive output targets. Spatial attention distillation uses the same anchors and positions.

Let f_θ and f_ϕ denote the student and teacher models, with parameters θ and ϕ, respectively. During on-policy training, the student generates a response trajectory from the original image and question:

$$y = (y_1, \ldots, y_T) \sim p_\theta(\cdot \mid I, x). \tag{1}$$

Here, T is the response length, and y_t is the token generated at position t. The teacher reuses the student trajectory and computes predictions under the same response prefix $y_{<t} = (y_1, \ldots, y_{t-1})$. Distillation supervision thus aligns with the decoding states that the student actually visits.

Figure 2 summarizes PA-OPSD in four stages, with shared privileged anchors linking position selection and attention supervision.

1. **Student on-policy sampling.** The student generates a response from the original image and question. This trajectory supplies the shared response prefixes for teacher and student.
2. **Privileged anchor extraction under multiple conditions.** The EMA teacher extracts complementary visual anchors from attention conditioned on the question, training-only answer, and early response. These anchors remain frozen in subsequent teacher passes.
3. **Critical generation position identification.** Under shared response prefixes and frozen anchors, the teacher compares predictions on the original image and a control image with the selected Top-K regions masked. Their JSD identifies visually sensitive generation positions.
4. **Dual-path selective distillation.** At the selected positions, the student matches contrastively enhanced output targets and the original-image teacher's normalized visual attention. The contrastive strength α controls target shaping, and both losses share the position mask.

3.2 Privileged Anchors from Multiple Conditions and Regions

We adapt the attention quality assessment and cross-layer aggregation ideas in ADAPT (Yao et al., 2026) to anchor extraction conditioned on the question, ground-truth answer, and student's early response. Our extensions aggregate the three conditions independently and fuse them adaptively. They explicitly retain multiple complementary evidence regions. The resulting frozen anchors support both critical generation position identification and visual attention distillation.

Throughout this paper, *visual attention* refers to attention from text tokens to image tokens. Let $\mathcal{V}$ be the visual token index set for image I, with $V = |\mathcal{V}|$ tokens. We read teacher attention from selected full-attention layers $\mathcal{L}$ and heads $\mathcal{H}_l$ in layer l. For layer l, head h, conditioning token i, and visual position $v \in \mathcal{V}$, let B^c_{lhiv} denote the raw attention under condition $c \in \{q, g, e\}$. The conditions q, g, and e correspond to the question, ground-truth answer, and early response. Their token sets are $\mathcal{Q}$, $\mathcal{G}$, and $\mathcal{E}$, respectively. We define the early response set as $\mathcal{E} = \{y_t \mid 1 \le t \le \min(K_e, T)\}$, where K_e caps its length. These tokens come from the student rollout, but their attention is recomputed by the teacher in the privileged context.

The three conditions play complementary roles. **Question tokens** describe what to look for and typically cover objects and diagrams relevant to the task. **Ground-truth answer tokens** indicate which evidence must support the final answer. They help exclude regions that are visually salient but irrelevant to the answer. **Early response tokens** indicate the student's current reasoning path and adapt the anchors to the current rollout. In mathematical diagrams and geometry problems, a correct answer may require reading values, labels, and spatial relations in sequence. A single condition often covers only part of this evidence. Combining the three conditions can better recover the full evidence set.

We first normalize teacher visual attention to obtain the relative focus at each image position. Next, we separately aggregate attention from the question, ground-truth answer, and early response. The resulting three saliency maps describe spatial visual focus and are fused using quality weights. The conditions may focus on different locations, preserving complementary visual cues.

We then identify highly attended regions in each condition map and the fused map. Duplicate candidates are removed to obtain at most K complementary privileged anchors. Each anchor stores its region, internal attention distribution, and confidence for subsequent teacher guidance and attention-ranked region masking. Retaining candidates from individual condition maps helps preserve evidence unique to one condition that fusion might otherwise suppress. Appendix A details aggregation, region extraction, and quality filtering.

3.3 Anchor-Guided Visual Contrast and Selective Output Distillation

The regional anchors identify visual evidence for the teacher to attend to. Response positions, however, differ in their dependence on this evidence. We let the same teacher observe the original image and a control image with the selected visual regions masked. Comparing its predictions under the same response prefix identifies positions with greater visual dependence. Distillation is then concentrated at those positions.

First, we combine frozen soft anchors by region confidence to adjust the teacher's text-to-visual attention scores. A spatial bias corrects large deviations from the anchors, while a uniform bias on visual tokens compensates for substantial drops in total visual attention. Ground-truth answers thus provide training-time visual guidance without entering the teacher's prediction context as text.

Next, we construct the control image I^- by directly masking the selected Top-K visual regions. Here, Top-K specifies a percentage of image regions ranked by aggregated attention, with Top-30% as the default. Pixels outside the selected regions, image dimensions, and the visual token layout remain unchanged. Top-0% leaves the control image unmasked, whereas Top-100% masks the entire image. This contrast measures how teacher predictions change when the selected visual evidence is hidden.

We refer to the teacher forward passes on the original image I and control image I^- as the original-image teacher (Positive Teacher) and the control teacher (Control Teacher). They share parameters, question x, student response prefix $y_{<t}$, and frozen anchors. Neither directly accesses the ground-truth answer. Let p_t^+ and p_t^- denote their prediction distributions at position t. Applying softmax to their output logits at the same temperature τ yields $\widehat{p_t^+}$ and $\widehat{p_t^-}$, respectively. We use Jensen–Shannon divergence (JSD) to measure prediction changes caused by masking the selected visual evidence:

$$s_t = \mathrm{JSD}(\widehat{p_t^+} \| \widehat{p_t^-}), \tag{2}$$

A larger s_t indicates greater sensitivity to the selected visual evidence. Let Ω denote the set of valid response prediction positions. We rank these positions by s_t in descending order and retain the top

$\lceil \rho|\Omega| \rceil$ candidates. Candidates below the sensitivity threshold τ_s are then removed. The parameter ρ controls the retained fraction. Selected positions receive $m_t = 1$; all others receive $m_t = 0$. Gradients are stopped through this mask. Output distillation and subsequent attention distillation share it, concentrating both supervision paths on the same critical visual decisions.

Following VCSD (Liang et al., 2026), we also use the visual contrast to enhance the output target. For each candidate token w in vocabulary $\mathcal{W}$, define $\Delta_t(w) = \log \widehat{p}_t^{+}(w) - \log \widehat{p}_t^{-}(w)$. The contrastively enhanced target is

$$q_t^{\alpha}(w) = \frac{\widehat{p}_t^{+}(w) \exp(\alpha \Delta_t(w))}{\sum_{u \in \mathcal{W}} \widehat{p}_t^{+}(u) \exp(\alpha \Delta_t(u))}. \tag{3}$$

Here, $\alpha \geq 0$ scales the log-probability contrast. Setting $\alpha = 0$ recovers the original-image teacher distribution. Normalization spans the full vocabulary without a plausibility threshold, and termination tokens receive the same contrastive adjustment. The JSD-based position mask remains computed from $\widehat{p}_t^{+}$ and $\widehat{p}_t^{-}$ before target shaping.

The student receives the original image, question, and shared response prefix, producing $\widehat{p}_t^{S}$ at the same temperature. At selected positions, it matches q_t^{α}:

$$\mathcal{L}_{\text{out}} = \frac{\tau^2}{N_m} \sum_{t \in \Omega} m_t D_{\text{KL}} \left(\text{sg}[q_t^{\alpha}] \| \widehat{p}_t^{S} \right), \qquad N_m = \max \left(1, \sum_{t \in \Omega} m_t \right). \tag{4}$$

Here, sg denotes stop-gradient, N_m normalizes by the selected-position count, and τ^2 compensates for temperature scaling. The control teacher contributes to both position selection and target shaping. The original-image teacher provides the base distribution and the visual attention target. Unselected positions produce no output distillation gradients.

This objective concentrates supervision on visually sensitive decisions while retaining a full-vocabulary target at each selected position. JSD determines where to supervise, while α controls how strongly visual contrast changes the output target.

3.4 Spatial Attention Distillation at Critical Positions

Similar output distributions do not imply that the student and teacher use the same visual evidence. We therefore reuse the position mask from Section 3.3 and supervise only generation positions with $m_t = 1$. The anchor-guided original-image teacher attention directly supervises the student's visual focus. Both models receive the original image and the same response prefix. Their visual tokens correspond in the same spatial coordinate system.

For the selected full-attention layers and heads, we first normalize text-to-visual attention over the full visual token set $\mathcal{V}$. We then aggregate it using weights over layers and heads. These weights are nonnegative, sum to one, and are shared by the teacher and student. This yields spatial probability distributions $\mathbf{a}_t^{T}$ and $\mathbf{a}_t^{S}$, where T denotes the original-image teacher and S the student. The selective spatial attention distillation loss is

$$\mathcal{L}_{\text{att}} = \frac{1}{N_m} \sum_{t \in \Omega} m_t \ \text{JSD} \left(\text{sg}[\mathbf{a}_t^{T}] \| \mathbf{a}_t^{S} \right). \tag{5}$$

Here, N_m uses the same selected-position normalization as output distillation. Gradients are stopped through both teacher distributions and aggregation weights. Normalization spans all visual tokens, rather than being performed separately within each anchor region. The loss therefore captures differences in both precise attention locations and relative allocation across regions. No separate regional losses are required. This objective constrains only the spatial distribution of visual attention; it does not explicitly match its total mass. By matching the anchor-guided teacher distribution, the student learns which visual evidence to attend to at critical generation positions. This complements prediction guidance at the output layer.

3.5 Overall Training Objective

On on-policy examples, PA-OPSD jointly optimizes selective output distillation and spatial attention distillation:

$$\mathcal{L}_{\mathrm{PA}} = \lambda_{\mathrm{out}}\mathcal{L}_{\mathrm{out}} + \lambda_{\mathrm{att}}\mathcal{L}_{\mathrm{att}}. \tag{6}$$

The coefficients λ_{out} and λ_{att} weight the two losses. The output loss uses q_t^{α} to guide predictions at critical positions; the attention loss transfers the original-image teacher's visual focus. Both paths share m_t and the privileged anchors, jointly aligning generation decisions and visual evidence allocation along student-generated trajectories. Appendix B presents the complete training procedure.

4 Experiments

4.1 Experimental Setup

We train Qwen3-VL-8B, Qwen3.5-4B, and Qwen3.5-9B on ViRL39K. Each teacher–student pair starts from the same checkpoint, with EMA teacher updates. Reference answers are used only for teacher anchor extraction. Evaluation covers MMMU (Yue et al., 2023), MMMU-Pro (Yue et al., 2024), RealWorldQA (RWQA) (xAI, 2024), MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2023), HR-Bench 4K/8K (HR4K/HR8K) (Wang et al., 2024), and HallusionBench (HalluB) (Guan et al., 2023). We report individual scores and their unweighted mean over eight settings, counting HR4K and HR8K separately. Training runs for one epoch (`max_steps=-1`) on 32 GPUs, with batch size 4 per device, one gradient accumulation step, and global batch size 128. We use a learning rate of 1×10^{-6}, warmup ratio 0.05, bf16, gradient checkpointing, and DeepSpeed ZeRO-2 for both models. Sequence and completion lengths are capped at 4096 and 2048, respectively. Defaults are Top-30% selection and $\alpha = 2.0$. For Qwen3.5, aggregation and attention distillation use only full-attention layers, with four selected layers for Qwen3.5-9B unless varied in the layer ablation.

Table 1: Results across three backbones. Avg. is the mean over eight benchmarks. Bold marks PA-OPSD scores exceeding all baselines within each backbone.

Base model	Method	MMMU	MMMU Pro	RWQA	Math Verse	Math Vista	HR4K	HR8K	HalluB	Avg.
Qwen3-VL-8B	Base	52.9	55.2	72.0	61.0	78.9	76.8	71.9	55.4	65.50
	OPSD	51.1	54.6	71.0	57.8	78.5	77.5	74.0	56.4	65.11
	VCSD	51.6	56.6	71.8	58.1	79.3	77.9	73.9	56.7	65.73
	RP-OPSD	53.7	55.5	71.2	58.1	78.6	78.3	74.2	57.0	65.83
	PA-OPSD	**54.0**	54.3	71.9	**62.1**	77.4	**79.0**	**75.8**	**58.2**	**66.58**
Qwen3.5-4B	Base	55.7	58.7	75.3	76.1	80.6	80.9	76.8	55.1	69.89
	OPSD	61.2	60.1	77.3	73.5	81.2	82.2	75.3	58.8	71.19
	VCSD	52.0	60.8	74.8	76.0	81.9	82.5	78.1	58.2	70.53
	RP-OPSD	56.7	60.1	75.3	74.0	81.5	82.9	75.0	57.6	70.39
	PA-OPSD	57.0	**60.8**	**79.5**	75.2	80.7	**85.3**	**78.4**	58.2	**71.88**
Qwen3.5-9B	Base	57.2	57.5	76.9	76.3	82.3	82.9	79.5	59.6	71.52
	OPSD	55.1	59.8	78.3	75.1	82.6	83.0	79.9	57.1	71.36
	VCSD	59.4	60.2	79.4	77.5	83.6	86.9	81.5	58.9	73.42
	RP-OPSD	57.4	59.9	78.0	76.2	83.2	84.3	78.88	55.5	71.67
	PA-OPSD	59.3	**63.2**	**80.7**	**78.3**	**83.7**	85.6	**83.0**	**62.6**	**74.55**

4.2 Main Results

Table 1 shows that PA-OPSD achieves the highest average on all three backbones, outperforming Base, OPSD (Zhao et al., 2026a), VCSD (Liang et al., 2026), and RP-OPSD (Zhu et al., 2026a). Average gains are 1.08–3.03 points over Base and 0.69–1.13 over the strongest competing distillation method for each backbone. PA-OPSD leads on 14 of 24 model–benchmark combinations, with

one additional tie, and achieves the best HR8K score on every backbone. These results support its effectiveness across model sizes and multimodal tasks.

4.3 Ablation Study

Table 2: Component ablations on Qwen3.5-9B with four full-attention layers. "Cond." denotes an anchor-conditioning branch.

Variant	MMMU	MMMU Pro	RWQA	Math Verse	Math Vista	HR4K	HR8K	HalluB
PA-OPSD	59.3	63.2	80.7	78.3	83.7	85.6	83.0	62.6
Random regions	56.3	62.0	80.0	77.8	84.3	85.0	82.6	61.3
w/o Early-Response Cond.	57.4	63.2	78.8	78.2	83.7	86.3	82.3	62.4
w/o Question Cond.	58.1	63.0	79.5	77.9	84.3	86.0	82.9	60.7
w/o Ground-Truth Cond.	58.0	62.9	78.6	78.0	84.5	84.6	82.9	59.9
w/o attn loss	58.4	63.6	79.4	78.5	83.2	84.0	83.3	61.2

Complementary conditioning cues. Removing early-response, question, or ground-truth conditioning lowers the average to 74.04, 74.05, and 73.68, respectively: drops of 0.51, 0.50, and 0.87 points. All three also reduce RWQA and MathVerse scores. This supports their complementary roles (Section 3.2), with the largest average drop occurring without the training-only ground-truth cue.

Anchor localization and spatial supervision. Replacing attention-derived regions with random regions, while retaining downstream procedures, lowers the average by 0.88 points to 73.67. MMMU and MMMU-Pro drop by 3.0 and 1.2 points. Removing spatial attention distillation lowers the average by 0.61 points to 73.94, including a 1.6-point HR4K drop. These results support both anchor localization and attention supervision beyond output matching; individual benchmark rankings vary.

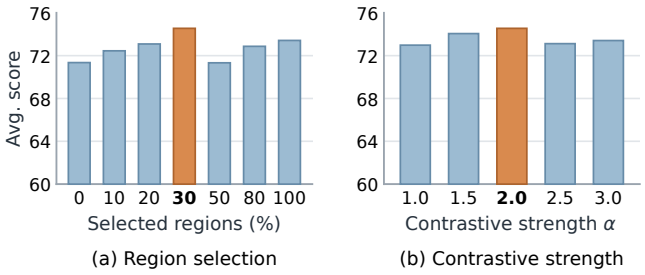

Figure 3: Qwen3.5-9B ablations: mean over eight evaluation settings.

Table 3: Masking sensitivity: original accuracy (%) and drops (pp).

		ΔAcc (pp)		
Dataset	Orig.	Top	Bot.	Rand.
MMBench	89.58	**5.31**	0.76	1.74
RWQA	75.82	**18.30**	0.26	3.27
VisNumBench	57.29	**8.73**	2.46	4.34

Region selection and contrastive strength. In Figure 3(a), Top-30% reaches 74.55, exceeding no masking by 3.20 points and full-image masking by 1.13 points. Panel (b) peaks at $\alpha = 2.0$, exceeding $\alpha = 1.0$ by 1.57 points. Larger values reduce performance. The sweeps fix $\alpha = 2.0$ and Top-30%, respectively, supporting targeted masking with moderate contrastive strength.

Number of full-attention layers. Table 4 varies the selected full-attention layers for aggregation and distillation. Four layers achieve the best average of 74.55, exceeding two, six, and eight layers by 0.19, 1.47, and 0.86 points, respectively. More layers do not consistently help, motivating the four-layer default. Individual benchmarks favor different settings.

Table 4: Full-attention layer-count ablation on Qwen3.5-9B. Bold marks the best average.

Layers	MMMU	MMMU Pro	RWQA	Math Verse	Math Vista	HR4K	HR8K	HalluB	Avg.
2	59.0	64.6	79.7	78.7	84.2	85.0	83.3	60.4	74.36
4	59.3	63.2	80.7	78.3	83.7	85.6	83.0	62.6	**74.55**
6	59.2	63.4	73.4	78.5	83.2	84.7	82.1	60.1	73.08
8	59.2	63.1	77.7	78.7	83.9	85.3	81.9	59.7	73.69

4.4 Inference-Time Validation of Anchor Localization

We test prompt-conditioned localization on MMBench (Liu et al., 2023), RealWorldQA (xAI, 2024), and VisNumBench (Weng et al., 2025), using non-thinking multiple-choice next-token scoring without training or logit adjustment. Attention is aggregated primarily over the last 128 prompt tokens. We mask equal-area regions with highest scores (Top), lowest scores (Bottom), or random locations (Random). Table 3 reports original accuracy and its drop after masking (ΔAcc).

Top masking causes the largest accuracy drop on all three benchmarks. On RealWorldQA, it produces an 18.30-point drop, versus 3.27 for Random and 0.26 for Bottom. Under equal-area masking, this greater disruption supports localization of prediction-relevant evidence, providing evidence for PA-OPSD's anchor-guided visual contrast and teacher attention guidance.

4.5 Qualitative Visualization of Conditional Attention

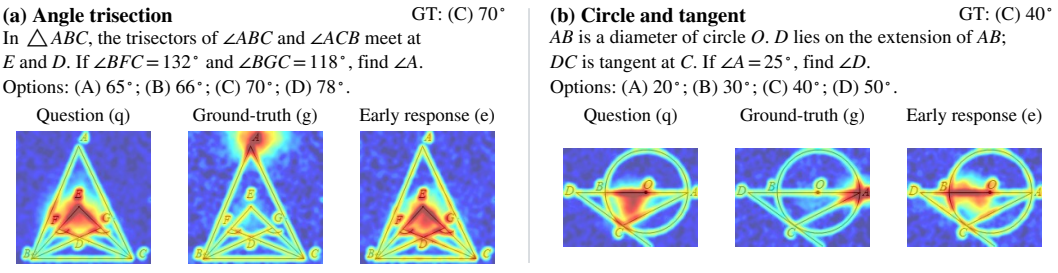

Figure 4: Conditional attention for two geometry examples. Each panel shows question (q), ground-truth answer (g), and early-response (e) maps; GT denotes the supplied answer.

Figure 4 compares the three conditioning branches on two geometry examples. In the angle-trisection example, the question and early-response maps emphasize internal intersections, while the ground-truth map peaks near vertex A. In the circle example, the maps highlight different parts of the diameter, interior, and endpoint A. These complementary patterns illustrate why PA-OPSD retains condition-specific maps before fusion.

5 Conclusion

We introduce PA-OPSD, which leverages multi-conditioned privileged visual anchors to concurrently identify visually sensitive decoding steps and align internal spatial attention along on-policy rollouts. PA-OPSD substantially improves visual reasoning while preserving inference architecture, offering a principled direction for grounding evidence in multimodal self-distillation.

AI use statement

Generative AI assisted with language editing, translation, methodological and experimental feedback, and result interpretation. The assisted content was checked for consistency with the method and experimental records. The authors take responsibility for the final manuscript.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. *arXiv preprint arXiv:2306.13649*, 2023. URL https://arxiv.org/abs/2306.13649.

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. *arXiv preprint arXiv:2608.05131*, 2026. URL https://arxiv.org/abs/2608.05131.

Zihan Chen, Hengguang Zhou, Yuan Kang, Yiming Zhang, Wenhui Fang, Zenghui Ding, Yining Sun, and Cho-Jui Hsieh. BAS-OPD: Budget-Aware Selective On-Policy Self-Distillation for Fine-Grained Multimodal Perception. *arXiv preprint arXiv:2609.25891*, 2026. URL https://arxiv.org/abs/2609.25891.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. *arXiv preprint arXiv:2310.14566*, 2023. URL https://arxiv.org/abs/2310.14566.

Jingtai He, Shiyuan Meng, Wenchao Meng, and Qinmin Yang. ADOPD: Reference-Privileged On-Policy Distillation for MLLM-Based Industrial Anomaly Detection. *arXiv preprint arXiv:2608.09789*, 2026. URL https://arxiv.org/abs/2608.09789.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Zesheng Yang, Fangzhi Xu, and Jun Liu. Visual-OPSD: Cross-Modal On-Policy Self-Distillation for Efficient Unified Multimodal Reasoning. *arXiv preprint arXiv:2606.18974*, 2026a. URL https://arxiv.org/abs/2606.18974.

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. *arXiv preprint arXiv:2608.14144*, 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. *arXiv preprint arXiv:2607.21556*, 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. *arXiv preprint arXiv:2605.21924*, 2026. URL https://arxiv.org/abs/2605.21924.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*, 2023. URL https://arxiv.org/abs/2307.06281.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. *arXiv preprint arXiv:2310.02255*, 2023. URL https://arxiv.org/abs/2310.02255.

Yunhang Qian, Jiaquan Yu, Jiawei Liu, Meng Wang, Hongwei Bran Li, and Xiaobin Hu. Med-OPD: Improving Medical Vision-Language Models via Evidence-Aware On-Policy Distillation. *arXiv preprint arXiv:2607.16303*, 2026. URL https://arxiv.org/abs/2607.16303.

Chenhao Qiu, Dawei Li, Yechao Zhang, Lei Gong, and Zhen Tan. OPD-Aha: From Linguistic Momentum to Visual Reflection in Multimodal On-Policy Distillation. *arXiv preprint arXiv:2609.16459*, 2026. URL https://arxiv.org/abs/2609.16459.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. ViCuR: Visual Cues as Recoverable Privilege for Multimodal On-Policy Distillation. *arXiv preprint arXiv:2606.05718*, 2026. URL https://arxiv.org/abs/2606.05718.

Shravan Venkatraman, Omkar Thawakar, Ritesh Thawkar, Abdelrahman Shaker, and Rao Muhammad Anwer. Perception Before Supervision: Self-Contained Visual Distillation from Counterfactual Blind Spots. *arXiv preprint arXiv:2608.09931*, 2026. URL https://arxiv.org/abs/2608.09931.

Kaishen Wang, Dongdi Zhao, Yijun Liang, Dingqiang Ye, Ruibo Chen, Heng Huang, and Di Fu. Where to Look Matters: On-Policy Self-Distillation for Long-Video Understanding. *arXiv preprint arXiv:2608.25356*, 2026a. URL https://arxiv.org/abs/2608.25356.

Shuai Wang, Daoan Zhang, Zhe Tang, Hao Cheng, and Jiaheng Wei. Self-Boosting Vision-Language Models with Noisy Student On-Policy Self-Distillation. *arXiv preprint arXiv:2607.23125*, 2026b. URL https://arxiv.org/abs/2607.23125.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, Conquer and Combine: A Training-Free Framework for High-Resolution Image Perception in Multimodal Large Language Models. *arXiv preprint arXiv:2408.15556*, 2024. URL https://arxiv.org/abs/2408.15556.

Ziyue Wang, Shiqi Huang, Weiwen Xu, Bihan Wen, and Xudong Jiang. Video-OPSD: Exploiting Privileged Visual Evidence for On-Policy Self-Distillation in Video Large Language Models. *arXiv preprint arXiv:2608.27065*, 2026c. URL https://arxiv.org/abs/2608.27065.

Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. VisNumBench: Evaluating Number Sense of Multimodal Large Language Models. *arXiv preprint arXiv:2503.14939*, 2025. URL https://arxiv.org/abs/2503.14939.

xAI. RealWorldQA. Dataset, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Leyan Xue, Feng Xiong, Mingjun Ma, and Changqing Zhang. Distill What the Student Can See: Fisher-Projected On-Policy Distillation for Vision-Language Models. *arXiv preprint arXiv:2608.01263*, 2026. URL https://arxiv.org/abs/2608.01263.

Zhiyuan Yao, Zheren Fu, Zhixiao Zheng, Jiajun Li, Yi Tu, and Zhendong Mao. ADAPT: Attention Dynamics Alignment with Preference Tuning for Faithful MLLMs. *arXiv preprint arXiv:2606.31054*, 2026. URL https://arxiv.org/abs/2606.31054.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-OPD: Confidence Aware Heterogeneous Multi-Teacher Multimodal On-policy Distillation. *arXiv preprint arXiv:2607.02592*, 2026. URL https://arxiv.org/abs/2607.02592.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to See Fine Details for Multimodal LLMs via On-Policy Self-Distillation. *arXiv preprint arXiv:2605.18740*, 2026. URL https://arxiv.org/abs/2605.18740.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502*, 2023. URL https://arxiv.org/abs/2311.16502.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. *arXiv preprint arXiv:2409.02813*, 2024. URL https://arxiv.org/abs/2409.02813.

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. *arXiv preprint arXiv:2607.28590*, 2026a. URL https://arxiv.org/abs/2607.28590.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? *arXiv preprint arXiv:2403.14624*, 2024. URL https://arxiv.org/abs/2403.14624.

Yan Zhang, Daiqing Wu, Huawen Shen, Can Ma, and Yu Zhou. Learn where to Click from Yourself: On-Policy Self-Distillation for GUI Grounding. *arXiv preprint arXiv:2605.00642*, 2026b. URL https://arxiv.org/abs/2605.00642.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. *arXiv preprint arXiv:2601.18734*, 2026a. URL https://arxiv.org/abs/2601.18734.

Tong Zhao, Yuyang Hu, Yutao Zhu, Reed Li, Haijin Liang, Haibo Shi, Yu Lu, and Zhicheng Dou. OPOD: On-Policy Omni Distillation. *arXiv preprint arXiv:2607.20918*, 2026b. URL https://arxiv.org/abs/2607.20918.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. RP-OPSD: Resolution-Privileged On-Policy Self-Distillation for Multimodal Large Language Models. *arXiv preprint arXiv:2607.24447*, 2026a. URL https://arxiv.org/abs/2607.24447.

Shoutai Zhu, Tianyang Xu, Bin Sun, Mingyuan Xu, Yu Liu, and Qinzhen Guo. OPLD: On-Policy Latent Distillation for Multimodal Reasoning. *arXiv preprint arXiv:2607.28154*, 2026b. URL https://arxiv.org/abs/2607.28154.

A Details of Privileged Anchor Construction from Multiple Conditions and Regions

This appendix details the attention quality assessment, condition fusion, and region extraction introduced in Section 3.2. We follow the notation in the main text.

For a decoder-only MLLM, text-to-visual attention is the submatrix whose queries are text tokens and whose keys are image tokens. It does not require a separate cross-attention module. We recover spatial locations from each example's visual token indices and image-grid metadata, rather than assuming a fixed image resolution or token count. For Qwen3.5, only full-attention layers provide these explicit attention matrices. The Qwen3.5-9B configuration uses four such layers, as specified in Section 4.1.

A.1 Conditional Attention and Quality Weights

Let $\mathcal{C}_q = \mathcal{Q}$, $\mathcal{C}_g = \mathcal{G}$, and $\mathcal{C}_e = \mathcal{E}$. During the anchor-extraction forward pass, the teacher reads the question, ground-truth answer, and student's early response. It separately records attention B^c_{lhiv} from the three types of conditioning tokens to visual tokens. Each condition uses an independent query mask. Irrelevant text, such as system prompts and template symbols, is excluded from aggregation.

For each valid query, we compute both the visual attention mass and the spatial distribution normalized over all visual tokens:

$$M^c_{lhi} = \sum_{v \in \mathcal{V}} B^c_{lhiv}, \qquad \bar{B}^c_{lhiv} = \frac{B^c_{lhiv}}{M^c_{lhi} + \epsilon}, \tag{A.1}$$

where $\epsilon > 0$ provides numerical stability. Retaining M^c_{lhi} distinguishes sufficient attention to the image from a concentrated distribution formed with very little visual attention.

Let $\mathbf{u}^c_{lh} = |\mathcal{C}_c|^{-1} \sum_{i \in \mathcal{C}_c} \bar{\mathbf{B}}^c_{lhi}$ be the mean visual distribution for a condition within the same layer and head. The quality weight of each query is

$$w^c_{lhi} = \left[\min\left(1, \frac{M^c_{lhi}}{\tau_m}\right) \left(1 - \frac{H(\bar{\mathbf{B}}^c_{lhi})}{\log V}\right) \left(1 - \frac{\mathrm{JSD}(\bar{\mathbf{B}}^c_{lhi} \| \mathbf{u}^c_{lh})}{\log 2}\right) \right]^{\gamma}. \tag{A.2}$$

The first term measures visual attention mass, with reference value $\tau_m > 0$. The second measures spatial concentration through normalized entropy. The third measures agreement between the current query and the mean distribution under the same condition. It reduces the influence of individual queries with anomalous shifts. Here, H is entropy, JSD denotes Jensen–Shannon divergence, and $\gamma > 0$ controls the strength of quality filtering. All three terms determine the aggregation weight. Concentration alone is not treated as a guarantee of correct evidence. Consistency is computed within each condition. The question, answer, and early response are not required to attend to the same region.

A.2 Within-Condition Aggregation and Cross-Condition Fusion

First, we aggregate visual distributions across selected layers, heads, and queries within each condition to obtain a conditional saliency map S^c:

$$S^c(v) = \frac{\sum_{l \in \mathcal{L}} \sum_{h \in \mathcal{H}_l} \sum_{i \in \mathcal{C}_c} w^c_{lhi} \bar{B}^c_{lhiv}}{\sum_{l \in \mathcal{L}} \sum_{h \in \mathcal{H}_l} \sum_{i \in \mathcal{C}_c} w^c_{lhi} + \epsilon}. \tag{A.3}$$

Condition quality is the mean of the corresponding query weights:

$$Q_c = \frac{\sum_{l \in \mathcal{L}} \sum_{h \in \mathcal{H}_l} \sum_{i \in \mathcal{C}_c} w^c_{lhi}}{\sum_{l \in \mathcal{L}} |\mathcal{H}_l| \, |\mathcal{C}_c|}. \tag{A.4}$$

These means are computed for nonempty sets of valid conditioning queries. If a condition has no valid query, both S^c and Q_c are set to zero, and it does not participate in fusion. If no valid condition map is available, anchor-based supervision is disabled for that example. We then fuse the maps according to condition quality:

$$S(v) = \sum_{c \in \{q,g,e\}} \beta_c S^c(v), \qquad \beta_c = \frac{Q_c}{\sum_{c'} Q_{c'} + \epsilon}. \tag{A.5}$$

Here, β_c is the normalized condition-quality weight. Normalizing within each condition before fusion by mean quality prevents longer answers or early responses from receiving greater weight simply because they contain more tokens. We retain S^q, S^g, S^e after fusion for subsequent region extraction. This helps preserve evidence unique to one condition that weighted averaging might weaken.

A.3 Region Extraction and Anchor Representation

Region extraction uses all three condition maps and the fused map:

1. Restore $\{S^q, S^g, S^e, S\}$ to spatial maps according to the example's visual token grid, and lightly smooth these attention maps.
2. Identify local high-response centers in each map. Apply non-maximum suppression to remove candidates with substantial spatial overlap while retaining complementary locations.
3. Starting from each retained center, grow a connected region through adjacent high-response locations using adaptive thresholds. If regions overlap, assign overlapping locations according to candidate saliency. This avoids counting the same location more than once in region confidence and image-mask coverage.
4. Retain at most K valid regions. If fewer candidates are available, use only the extracted regions without adding irrelevant locations. If no valid region exists, skip anchor-based supervision for that example.

Let $R_k(v) \in \{0, 1\}$ denote the mask of region k. A value of one indicates that visual position v belongs to the region. The soft anchor within the region and its confidence are

$$a_k(v) = \frac{R_k(v)S(v)}{\sum_{v' \in \mathcal{V}} R_k(v')S(v') + \epsilon}, \qquad r_k = \sum_{v \in \mathcal{V}} R_k(v)S(v). \tag{A.6}$$

Here, $\mathbf{a}_k = (a_k(v))_{v \in \mathcal{V}}$ describes relative attention within the region. The confidence r_k is its total saliency mass. We retain $\mathcal{A} = \{(R_k, \mathbf{a}_k, r_k)\}_{k=1}^{K}$, iterating only over valid regions when fewer than K exist. The region masks locate visual evidence for attention-ranked Top-K masking of the control image. The soft anchor $\mathbf{a}_k$ provides spatial guidance within the region, and r_k weights teacher guidance. The masking extent is controlled by the selected region percentage.

The anchor-count cap K and the Top-K masking percentage are distinct quantities. The former limits the number of complementary anchor regions; the latter controls how much of the image is masked after attention ranking. Spatial smoothing above applies to saliency maps during localization. The control image itself uses direct masking, as specified in Section 3.3.

A.4 Example-Level Quality Gating and Frozen Reuse

The example-level anchor quality η combines visual quality across conditions with cross-layer consistency. It determines whether the localized regions are suitable for supervision, using threshold τ_a. Anchor-derived output and attention supervision is enabled only when $\eta \geq \tau_a$ and at least one valid region exists. Otherwise, $\mathcal{L}_{\mathrm{PA}}$ is set to zero for that example; other valid examples in the batch can still provide training signals.

This gate only determines whether the anchors are usable. It neither ranks response positions nor replaces the visual sensitivity selection in Section 3.3. After anchor construction, gradients are stopped

through region masks, soft anchors, confidence values, and example quality. These quantities remain fixed within the current training iteration. The original-image and control teachers reuse the same anchors. Regions are not re-extracted from the control image, preventing localization changes from confounding the visual contrast. Both output-position selection and spatial attention distillation use teacher signals guided by these fixed anchors. The two supervision paths thus share the same visual evidence.

B Training Procedure and Implementation Details

This appendix expands the teacher guidance, sequence alignment, and training procedure in Section 3. It uses the same two-loss objective in Eq. (6).

B.1 Shared Rollouts and Prediction-Position Alignment

The student first samples y from (I, x). The sampled token IDs are then fixed for the current iteration. The teacher evaluates this trajectory by teacher forcing, rather than generating another answer. Original-image teacher, control teacher, and student predictions at position t therefore condition on the same $y_{<t}$. A causal forward pass over the complete response can evaluate these prefixes in parallel without exposing future response tokens to the prediction of y_t.

The privileged anchor pass has a different text layout because it contains $a^\star$. Response spans must therefore be located from their semantic boundaries instead of matched by absolute sequence offsets. We align the response token IDs, next-token prediction positions, and visual indices across passes. In particular, the logits and attention query used for step t must correspond to predicting y_t, not to a state that has already consumed it. This also avoids offsets caused by templates or padding. The valid set Ω includes targets in the response and excludes targets belonging to the prompt or padding.

After anchor extraction, both teacher prediction contexts are rebuilt without the ground-truth answer. Only the frozen spatial anchors and their quality information are carried over; cached text states from the privileged pass must not enter these predictions. The student receives neither answer text nor anchors as additional inputs. Aligning the prefixes in this way prevents differences in language context from being mistaken for visual sensitivity.

B.2 Frozen Spatial Guidance for the Teacher

We combine valid regional anchors into a spatial prior using their confidence:

$$g(v) = \frac{\sum_k r_k a_k(v)}{\sum_k r_k + \epsilon}, \qquad b(v) = \log\big(g(v) + \epsilon\big) - \frac{1}{V}\sum_{v' \in \mathcal{V}} \log\big(g(v') + \epsilon\big). \tag{B.1}$$

The centered log prior b favors anchor locations while removing a uniform offset from the spatial bias. Let z_{lhtv} denote a teacher attention logit from the query for response prediction t to visual token v. The correction described in Section 3.3 can be written as

$$\widetilde{z}_{lhtv} = z_{lhtv} + \kappa_{\mathrm{spa}}\, \chi_{lht}\, b(v) + \kappa_{\mathrm{mass}}\, \delta_{lht}, \qquad v \in \mathcal{V}. \tag{B.2}$$

Here, $\chi_{lht} \in \{0, 1\}$ activates the spatial correction only when the current visual focus deviates sufficiently from the frozen anchors. The nonnegative term δ_{lht} is zero unless total visual attention has fallen sufficiently below its reference level. It applies the same restorative bias to every visual key. The nonnegative coefficients κ_{spa} and κ_{mass} control these teacher-side corrections; they are distinct from the student loss weights.

The corrections are applied before the attention softmax, without directly modifying text-key logits. The spatial term changes relative preferences among visual locations, whereas the uniform visual term adjusts visual participation relative to text. Centering the spatial term does not itself guarantee a fixed total visual attention mass after softmax. These operations guide the teacher; they do not introduce a visual-mass matching loss for the student.

Both teacher views reuse the same $\mathcal{A}$, spatial prior, guidance coefficients, and activation rules. Their current attention can differ because their images differ. Regions are never re-extracted from the masked image. The teacher is thus allowed to respond to missing evidence under a fixed spatial

reference, rather than changing the reference itself. The original-image pass supplies the spatial target because it retains the full visual evidence and shares the student's observable input.

B.3 Direct Masking and Shared Position Selection

For clarity, let $\kappa \in [0, 1]$ denote the Top-K image masking fraction, separate from the anchor-count cap K and response-position fraction ρ. Attention-ranked regions define a binary mask $\widehat{R}_\kappa$ in the corresponding image or patch coordinates. If I_{mask} denotes the replacement content at masked locations, the control view is

$$I^- = (1 - \widehat{R}_\kappa) \odot I + \widehat{R}_\kappa \odot I_{\text{mask}}. \tag{B.3}$$

This operation leaves unselected pixels, image shape, visual token count, and spatial indices unchanged. Top-0% has $\widehat{R}_\kappa = 0$; Top-100% masks the full image. The default is Top-30%. No Gaussian blur or noise mixture is used to construct this view.

For probability distributions p and q on the same support, we use

$$\text{JSD}(p\|q) = \tfrac{1}{2} D_{\text{KL}}(p\|u) + \tfrac{1}{2} D_{\text{KL}}(q\|u), \qquad u = \tfrac{1}{2}(p + q). \tag{B.4}$$

With natural logarithms, this symmetric divergence lies in $[0, \log 2]$. Comparing the full vocabulary distributions in Eq. (2) captures changes in alternative predictions and uncertainty, even when the most likely token remains unchanged. The score measures sensitivity under the specified masking intervention; it is not an unconditional estimate of causal importance.

Let $\mathcal{T}_\rho$ be the $\lceil \rho|\Omega| \rceil$ valid positions with the largest s_t. The main-text selection rule is equivalently

$$m_t = \mathbf{1}\{t \in \mathcal{T}_\rho\}\, \mathbf{1}\{s_t \geq \tau_s\}, \qquad t \in \Omega. \tag{B.5}$$

Ranking allocates supervision within each response. The absolute threshold prevents a fixed fraction from being selected when all visual changes are too small. Consequently, fewer than $\lceil \rho|\Omega| \rceil$ positions may survive, including none. Selection is based on teacher sensitivity, not a predefined list of token types; intermediate reasoning decisions can be selected alongside final-answer tokens.

The mask is formed before contrastive target shaping and is shared by the output and spatial attention losses. If the anchors fail the quality gate or no position survives, both losses contribute zero for that example. The definition $N_m = \max(1, \sum_t m_t)$ avoids division by zero. A valid termination-token prediction follows the same selection rule and receives the same contrastive adjustment as any other selected position.

B.4 Target Construction and Spatial Attention Aggregation

An equivalent form of Eq. (3) makes the role of α explicit:

$$q_t^\alpha = \text{Softmax}_{\mathcal{W}}\big((1 + \alpha) \log \widehat{p}_t^+ - \alpha \log \widehat{p}_t^-\big)\,. \tag{B.6}$$

Log-softmax and log-sum-exp allow this target to be evaluated without explicitly exponentiating large probability ratios. Normalization includes the entire vocabulary. The control distribution contributes to both s_t and q_t^α, while the original-image distribution supplies the base target. Setting $\alpha = 0$ recovers $\widehat{p}_t^+$; the default $\alpha = 2.0$ amplifies the log-probability difference. All target quantities are detached from the student optimization.

For attention distillation, let A^r_{lhtv}, with $r \in \{T, S\}$, be the original-image teacher or student attention at the query predicting y_t. We normalize each layer–head distribution over the complete visual token set, then aggregate:

$$\begin{aligned} \bar{A}^r_{lhtv} &= \frac{A^r_{lhtv}}{\sum_{v' \in \mathcal{V}} A^r_{lhtv'}}, & a^r_t(v) &= \sum_{l \in \mathcal{L}} \sum_{h \in \mathcal{H}_l} \omega_{lh} \bar{A}^r_{lhtv}, \\ \omega_{lh} &\geq 0, & \sum_{l \in \mathcal{L}} \sum_{h \in \mathcal{H}_l} \omega_{lh} &= 1. \end{aligned} \tag{B.7}$$

The normalization is defined for positive visual attention mass. Numerical implementations stabilize division and logarithms while retaining unit-normalized spatial distributions. The same nonnegative

weights and corresponding full-attention layers and heads are used for both models. The weights are treated as constants during backpropagation. Gradients pass through the student's attention and its normalization, but not through the teacher target.

Equation (5) applies one JSD to the resulting distributions $\mathbf{a}_t^T$ and $\mathbf{a}_t^S$ at each selected position. Because the normalization includes all visual tokens, shifting attention between separate regions changes this loss. Local shifts within a region also change it. The objective does not separately normalize each anchor, average per-region losses, or align absolute visual attention mass. Thus, the detailed aggregation retains exactly the spatial objective specified in the main text.

B.5 Gradient Flow, EMA Updates, and Memory Use

The anchor, original-image, and control teacher passes all run without gradients. Region masks, confidence values, quality gates, position masks, teacher probabilities, and teacher attention targets remain fixed during the student update. The sampled rollout is also treated as a fixed sequence; gradients are not propagated through discrete sampling or position ranking. Only the student parameters are optimized by Eq. (6).

Teacher and student start from the same checkpoint. After a student optimizer step, the teacher is updated as $\phi \leftarrow \mu\phi + (1-\mu)\theta$, with $\mu \in [0, 1)$. EMA smooths changes in the target model while allowing it to evolve with the student. The teacher remains fixed across all passes within an iteration. If an entire batch contains no eligible supervision, the iteration provides no PA-OPSD gradient and no student or EMA update is required.

For memory efficiency, teacher passes precede the gradient-enabled student pass. Intermediate teacher activations can be released after retaining the distributions and attention records needed by the losses. Attention extraction is restricted to the selected full-attention layers and heads, and the attention loss uses only queries selected by m_t. Uniform query sampling cannot replace the sensitivity mask. Long responses can be processed in position chunks, accumulating both loss sums before applying their common selected-position normalization. This preserves the full vocabulary at each position and the complete selected-position objective. Gradient checkpointing and the distributed settings remain as reported in Section 4.1.

The teacher guidance, privileged context, control image, and distillation computations are used only during training. Inference retains the ordinary student forward pass on the image and question.

B.6 Complete On-Policy Training Iteration

Algorithm 1 includes the quality and empty-selection checks. For a batch, the same steps are applied to its examples before the optimizer update.

Algorithm 1 One on-policy training iteration of PA-OPSD

Require: Example $(I, x, a^\star)$, student f_θ, EMA teacher f_ϕ, $\mu \in [0, 1)$, $\alpha \geq 0$
Ensure: Student and teacher parameters; unchanged if no valid supervision exists

Sample $y \sim p_\theta(\cdot \mid I, x)$; fix the sampled tokens for this iteration
Align response token IDs, prediction queries, and visual indices across all passes
Teacher computations below use fixed parameters and no gradients.
Extract attention conditioned on x, $a^\star$, and $y_{1:\min(K_e,T)}$
Aggregate quality-weighted attention using Eqs. (A.2)–(A.5)
Extract $\mathcal{A}$ and quality η; stop gradients through both
if $\eta < \tau_a$ or no valid anchor exists **then**
    $\mathcal{L}_{\mathrm{PA}} \leftarrow 0$; **return** θ, ϕ unchanged
end if
Rebuild teacher prediction contexts without $a^\star$ or privileged cached text states
Form the frozen spatial prior and teacher guidance using Eqs. (B.1)–(B.2)
Run the guided original-image teacher on (I, x, y); retain $\{\widehat{p}_t^{+}\}$ and attention
Directly mask the selected Top-K regions to form I^-, preserving visual indices
Run the control teacher on (I^-, x, y) with the same anchors and guidance rules
Compute $s_t = \mathrm{JSD}(\widehat{p}_t^{+} \| \widehat{p}_t^{-})$ before target shaping
Select m_t with Eq. (B.5); let $n_m = \sum_{t \in \Omega} m_t$
if $n_m = 0$ **then**
    $\mathcal{L}_{\mathrm{PA}} \leftarrow 0$; **return** θ, ϕ unchanged
end if
Form and detach full-vocabulary targets q_t^α using Eq. (3)
Student computation below enables gradients.
Run the student on (I, x, y) to obtain $\{\widehat{p}_t^{S}\}$ and visual attention
Aggregate spatial attention with shared detached weights using Eq. (B.7)
Compute $\mathcal{L}_{\mathrm{out}}$ and $\mathcal{L}_{\mathrm{att}}$ with the same m_t and N_m
Optimize θ with $\mathcal{L}_{\mathrm{PA}} = \lambda_{\mathrm{out}}\mathcal{L}_{\mathrm{out}} + \lambda_{\mathrm{att}}\mathcal{L}_{\mathrm{att}}$
After the student update, set $\phi \leftarrow \mu\phi + (1 - \mu)\theta$