VISUAL CAUSAL TRANSPORT DISTILLATION VIA COUNTERFACTUAL RELATIVE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual on-policy self-distillation transfers privileged visual information mainly through a teacher's output predictions, leaving how visual evidence is used implicit. In a training-set diagnostic, image-removed accuracy rises from 35.8% to 41.6% after Vision-OPD, suggesting that part of the acquired knowledge is accessible through image-independent priors. We introduce Visual Causal Transport Distillation (VCT), which augments output distillation with counterfactual relative alignment of hidden representations. For each student-generated prefix, an EMA teacher is evaluated with privileged visual input and with its visual embeddings zeroed. Their contrast provides a reference for aligning the student's final representation and cross-depth updates toward evidence-conditioned computation. A dimensionless distance ratio performs relative alignment, while intervention energy and bounded evidence-gap weights focus supervision on visually sensitive tokens and trajectories. Across six benchmarks, VCT achieves average accuracies of 72.10% and 75.11% with Qwen3.5-4B and Qwen3.5-9B, outperforming Vision-OPD by 1.39 and 3.57 percentage points, respectively. VCT adds only one intervened teacher pass during training and leaves student inference unchanged.

1 INTRODUCTION

Multimodal language models can answer correctly by inspecting an image, recalling a training association, or exploiting regularities in questions and answer choices. Accuracy alone does not distinguish these routes. A teacher with privileged visual evidence can improve a student's predictions, yet matching those predictions does not explicitly specify which evidence-dependent computations the student should reproduce.

On-policy self-distillation supervises student-generated trajectories using a teacher with privileged context (Zhao et al., 2026; Cheng et al., 2026). Visual variants provide localized evidence, higher

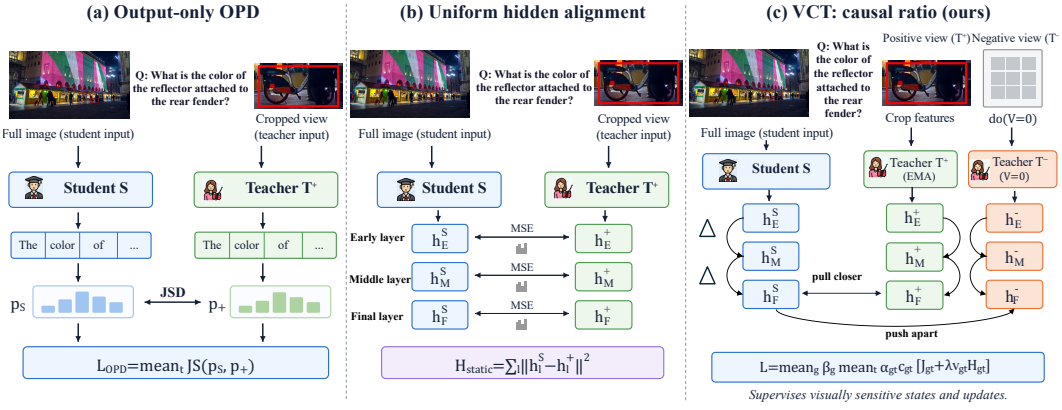

Figure 1: Comparison of other OPD mechanisms. Output-only OPD matches predictions; uniform hidden alignment regresses teacher states. VCT introduces a matched visual intervention and relative-distance supervision over final states and cross-depth updates.

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.

resolution, or cleaner inputs to the teacher (Yuan et al., 2026; Zhu et al., 2026; Li et al., 2026b), while region-to-image distillation transfers answers from informative crops to full-image inputs (Wei et al., 2026). Although their objectives differ, these approaches transfer privileged information through output supervision. Figure 1 contrasts this interface with uniform hidden alignment and our intervention-conditioned alternative.

Our motivation comes from an image-removed diagnostic: after Vision-OPD training, accuracy on the training questions rises from 35.8% to 41.6% without image input. Some acquired answer information is therefore accessible without inspecting the image at inference. This observation is consistent with stronger language priors or memorization; it does not establish deteriorating visual grounding. It instead motivates the question: *can distillation supervise hidden computations associated with visual evidence, in addition to their output predictions?*

We introduce Visual Causal Transport Distillation (VCT) to address this question. For each student-generated prefix, the same teacher is evaluated with privileged visual embeddings and with those embeddings zeroed, preserving its text and token layout. This controlled intervention defines a reference for the teacher's dependence on visual input. We use it to supervise both the final hidden representation and representation updates across depth.

The central objective is *counterfactual relative alignment*: the student minimizes its distance to the positive teacher relative to its combined distances to the positive and intervened teachers. Unlike an absolute squared error, this dimensionless ratio measures progress against an evidence-absent reference. Teacher intervention energies select sensitive hidden objects, while positive–negative and student–teacher output divergences jointly determine bounded token and rollout weights. Per-rollout averaging prevents supervision from growing automatically with response length.

Our contributions are as follows:

- We introduce **Visual Causal Transport Distillation (VCT)**, an on-policy self-distillation framework that constructs a counterfactual teacher reference by intervening on visual embeddings. This design extends privileged-information transfer beyond output predictions to hidden representations and cross-depth representation updates.
- We propose **counterfactual relative alignment** together with **bounded evidence-gap allocation**. The relative objective aligns the student toward the evidence-conditioned teacher with respect to an evidence-absent reference, while intervention energy and output divergence identify visually sensitive representation objects, tokens, and trajectories on which to concentrate supervision. We further characterize the reference-relative geometry of the resulting ratio objective.
- We evaluate VCT on six multimodal benchmarks with Qwen3.5-4B and Qwen3.5-9B. VCT achieves the highest average accuracy among the compared methods at both model scales, improving over Vision-OPD by **1.39** and **3.57** percentage points, respectively. It obtains the best reported result on all six tasks at 4B and on five of six tasks at 9B, while tying the best result on TreeBench.

2 Related Work

Distillation and representation transfer. Knowledge distillation transfers teacher distributions or generated sequences (Hinton et al., 2015; Kim & Rush, 2016). Same-architecture and within-network self-distillation show that improvement need not require a larger teacher (Furlanello et al., 2018; Zhang et al., 2019), with further work examining regularization and target decomposition (Mobahi et al., 2020; Zhao et al., 2022). Beyond outputs, feature hints, attention transfer, and cross-stage review supervise intermediate computations (Romero et al., 2015; Zagoruyko & Komodakis, 2017; Chen et al., 2021); relational and contrastive objectives preserve representation structure (Park et al., 2019; Tian et al., 2020). Moving-average targets and stop-gradient are also central to self-supervised representation learning (Tarvainen & Valpola, 2017; Grill et al., 2020; Caron et al., 2021; He et al., 2020; Chen & He, 2021). VCT constructs its hidden reference from the same teacher and prefix under a visual intervention, rather than an unrelated example or student augmentation.

On-policy distillation and privileged context. Learner-induced prefixes are a central concern in sequential imitation (Ross et al., 2011). Language-model distillation explores off-policy demon-

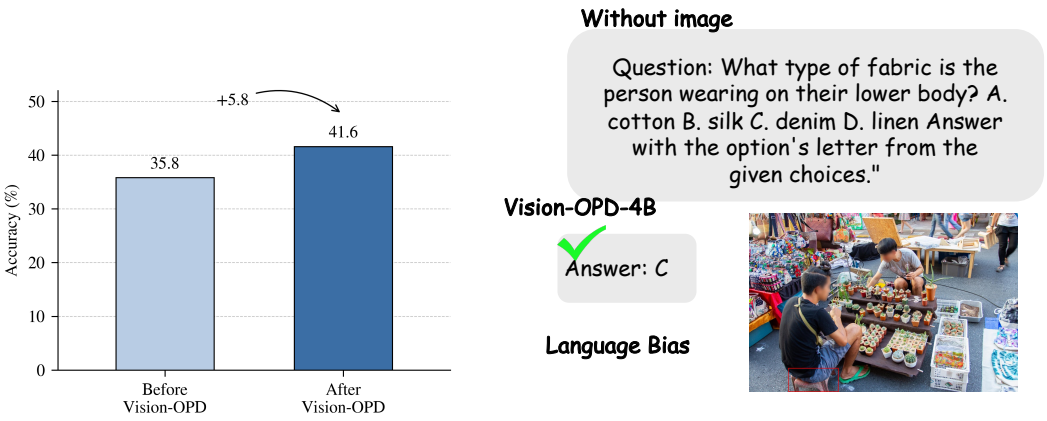

Figure 2: Empirical motivation: answers remain accessible without the image. Left: image-removed training-set accuracy increases from 35.8% to 41.6% after Vision-OPD (+5.8 points). Right: an author-supplied Vision-OPD-4B example answers the fabric question correctly without image input.

strations, student-generated trajectories, and alternative divergence objectives (Pang & He, 2021; Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024; 2025), including entropy-aware supervision (Jin et al., 2026). Context distillation internalizes behavior elicited by additional information (Snell et al., 2022); self-play and self-rewarding approaches exploit model-generated feedback (Chen et al., 2024b; Yuan et al., 2024). Recent self-distillation methods use privileged context, feedback, sampled solutions, or rationales to supervise on-policy responses (Zhao et al., 2026; Hübotter et al., 2026; Li et al., 2026c; Cheng et al., 2026). We retain student-prefix replay, but use the teacher's visual intervention response to define both hidden references and supervision priorities.

Visual evidence in multimodal learning. Vision-OPD and region-to-image distillation transfer localized evidence through output supervision (Yuan et al., 2026; Wei et al., 2026). Resolution privilege, asymmetric augmentation, and temporal feedback extend multimodal distillation (Zhu et al., 2026; Li et al., 2026b;a). Contrasting visual conditions has also informed decoding and distillation (Leng et al., 2024; Sun et al., 2026; Zhang et al., 2026). Our distinction is a hidden relative-distance objective over final states and cross-depth updates: the intervened teacher supplies a geometric reference as well as an allocation signal.

3 Motivation

Figure 2 shows that, after Vision-OPD training, accuracy on the same training questions increases from 35.8% to 41.6% even when the image is removed, corresponding to a 5.8-point gain. This result suggests that a non-negligible portion of the transferred knowledge can be accessed without visual evidence at inference time. In other words, existing privileged-information distillation may encourage the student to absorb visual–textual associations, answer regularities, or other dataset-level priors that correlate with the teacher's predictions, rather than explicitly learning the evidence-dependent computation by which the teacher interprets the image. Such priors can still improve answer accuracy, but they do not specify *how* visual information should affect the student's internal reasoning process.

This observation motivates us to move beyond answer transfer and explicitly distill the teacher's *visual dependence*. Our goal is not merely to make the student reproduce what the teacher predicts, but to encourage it to reproduce how the teacher's internal representations change in response to visual evidence. To this end, we construct a controlled counterfactual intervention on the teacher: for the same sampled prefix, we compare its computation with privileged visual evidence against a matched computation in which the visual embeddings are removed. The resulting representation difference provides a model-internal direction of visual influence. By aligning the student's representations and cross-depth updates with the evidence-conditioned teacher relative to this counterfactual reference, we encourage the student to follow the teacher's visual processing trajectory—that is, to learn not only

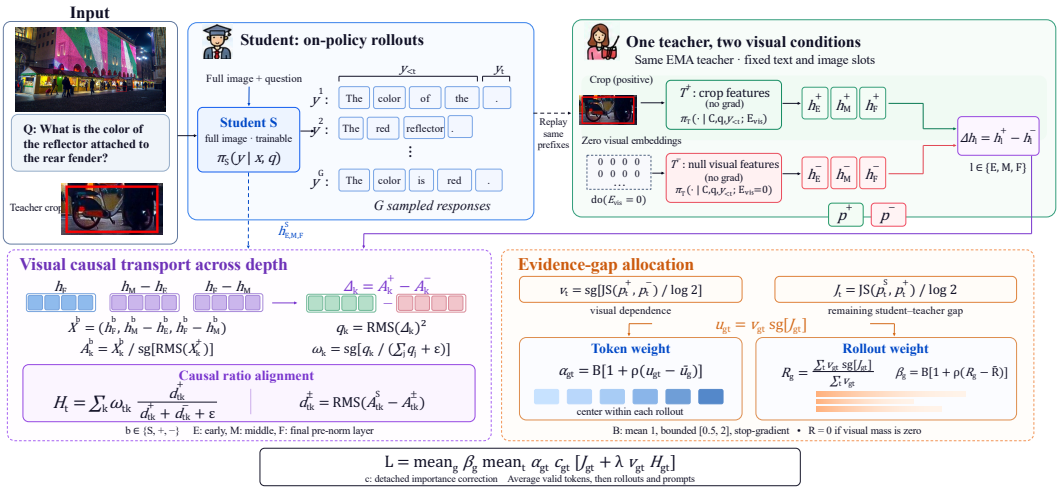

Figure 3: Overview of VCT. The full-image student generates on-policy trajectories. The EMA teacher replays each prefix with privileged crop embeddings and with those embeddings zeroed. Teacher intervention energies weight the final-state and cross-depth objects; a relative-distance loss aligns the student with the positive teacher. Evidence-gap scores determine bounded token and rollout weights for both objectives. The prompt index and numerical validity masks are omitted for clarity; c is a detached importance correction, independent of the allocation strength ρ. Bars are schematic.

the teacher's answer, but also a representation-level pattern of *how to use the image* when producing that answer.

4 Method

4.1 On-policy replay with one teacher intervention

Let x be a full image, q its question, and x^+ a privileged region view supplied by the Vision-OPD data. For each of B prompts, the student π_θ samples G trajectories $y^{(g)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x, q)$. Training replays each trajectory rather than generating a separate teacher response. At prediction position t, the three branches are

$$p_t^S = \pi_\theta(\cdot \mid x, q, y_{<t}), \qquad p_t^+ = \pi_\phi(\cdot \mid C^+(x^+, q), y_{<t}),$$
$$p_t^- = \pi_\phi(\cdot \mid C^+(x^+, q), y_{<t}; \mathrm{do}(E_{\text{vis}} = 0)). \tag{1}$$

Here C^+ denotes the dataset's teacher prompt, which may differ from the student's prompt. The two teacher branches share exactly the same prompt, parameters, masks, visual-token positions, and positional layout. The intervention replaces projected visual embeddings at their input slots with zeros; it does not remove image tokens or replace image pixels with black pixels. The negative branch skips the vision encoder. Its later hidden states remain nonzero because text, positions, and transformer computation remain active.

The teacher is an exponential moving average (EMA) of the student, initialized from the same checkpoint. After an actor update, $\phi \leftarrow (1 - \tau)\phi + \tau\theta$, with $\tau = 0.05$. Both teacher branches run without gradients. Hidden states and logits are aligned at positions that *predict* the same response token y_t, despite potentially different prompt lengths. The model predicts every evaluated token from the common sampled prefix $y_{<t}$.

Output distillation and visual dependence. For computational efficiency, the implementation selects the student's top $K = 100$ vocabulary entries and appends a tail bucket containing all remaining probability mass. The same support is used for S, T^+, T^-. Let $\tilde{p}$ denote this coarsened

distribution and define

$$\overline{\mathrm{JS}}(p,q) = \frac{\mathrm{KL}(p\|m) + \mathrm{KL}(q\|m)}{2\log 2}, \quad m = \tfrac{1}{2}(p+q),$$
$$J_t = \overline{\mathrm{JS}}(\tilde{p}_t^S, \tilde{p}_t^+), \qquad v_t = \mathrm{sg}[\overline{\mathrm{JS}}(\tilde{p}_t^+, \tilde{p}_t^-)]. \tag{2}$$

Both quantities lie in $[0,1]$. J_t is the trainable output-distillation loss; v_t is a detached, intervention-based visual-dependence score. It measures how much the teacher's next-token distribution changes when visual embeddings are zeroed, conditional on the current prefix. It is not a hard visual/nonvisual token label. Because it uses a top-K partition, it can miss changes among vocabulary entries grouped into the tail.

4.2 Visual transport objects across depth

For a transformer with L blocks, we capture outputs at $\ell_E = \lceil 0.25L \rceil$, $\ell_M = \lceil 0.5L \rceil$, and $\ell_F = L$. These are block outputs *before the final model normalization*, including the final selected layer. For each branch $b \in \{S, +, -\}$ and response position, define

$$X_1^b = h_F^b, \qquad X_2^b = h_M^b - h_E^b, \qquad X_3^b = h_F^b - h_M^b. \tag{3}$$

The first object anchors the final representation. The other two represent updates across depth, preserving how the model transforms the prefix representation over two intervals. They are not differences between neighboring response tokens. No branch-specific unit-vector normalization is applied before forming these differences.

Different objects can have different magnitudes. We express all three branches of each object in common teacher-defined units. Writing $\mathrm{RMS}(z) = \|z\|_2/\sqrt{d}$ for hidden width d, define

$$s_k = \mathrm{sg}[\mathrm{RMS}(X_k^+)], \quad b_k = \mathbf{1}[s_k > \epsilon], \quad \bar{s}_k = \begin{cases} s_k, & b_k = 1, \\ 1, & b_k = 0, \end{cases} \qquad A_k^b = X_k^b / \bar{s}_k. \tag{4}$$

The scale and validity mask are detached. Unlike independently normalizing each vector, a common scale retains relative student–teacher amplitudes. Invalid objects receive zero loss weight.

The teacher intervention defines $\Delta_k = A_k^+ - A_k^-$. Its energy selects the representation objects most affected by visual input:

$$q_k = b_k \,\mathrm{RMS}(\Delta_k)^2, \qquad \omega_k = \mathrm{sg}\left[\frac{q_k}{\sum_{j=1}^3 q_j + \epsilon}\right]. \tag{5}$$

For the middle-depth update, this difference has the explicit form

$$\Delta_2 = \frac{(h_M^+ - h_M^-) - (h_E^+ - h_E^-)}{\bar{s}_2}. \tag{6}$$

Visual transport thus describes changes in the intervention effect across depth. The student aligns to teacher updates selected by this sensitivity; no negative student branch is required.

The final-state and update objects serve complementary roles. Matching only h_F constrains the endpoint but leaves the intermediate transformation unconstrained. Matching $h_M - h_E$ and $h_F - h_M$ also supervises how the representation changes over depth. Their positive–negative differences identify intervals in which the teacher's transformation responds to visual input. The shared teacher scale preserves relative amplitudes across branches, while ω_k allows the three objects to contribute differently at each response position without introducing separate layer coefficients.

4.3 Counterfactual relative alignment

For each object, compute positive and negative RMS distances in the shared units:

$$d_k^+ = \mathrm{RMS}(A_k^S - A_k^+), \qquad d_k^- = \mathrm{RMS}(A_k^S - A_k^-),$$
$$r_k = b_k \frac{d_k^+}{d_k^+ + d_k^- + \epsilon}, \qquad H_t = \sum_{k=1}^3 \omega_k r_k. \tag{7}$$

Gradients flow through both student distances; teacher references remain detached. The ratio measures closeness to the positive teacher against an evidence-absent reference, while ω_k emphasizes objects changed by the intervention. The total objective additionally gates H_t by v_t, requiring a corresponding change in the teacher's output distribution. Alignment applies to each selected vector as a whole; it does not isolate a pure visual subspace.

Both H_t and J_t are dimensionless and bounded by one. Their empirical means need not match, since $v_t H_t$ decreases with visual dependence and alignment error. We use a single balance parameter $\lambda = 1$, without adaptive loss-magnitude matching.

The negative reference also changes the geometry of supervision. When the two teacher references are well separated, a student equidistant from them receives a ratio near one half; approaching the positive teacher drives the ratio toward zero. Thus, the loss encodes relative preference for the evidence-conditioned representation. It does not encourage unconstrained separation from the negative: far from both fixed references, the ratio approaches one half. Output distillation provides predictive supervision alongside this hidden geometric constraint.

4.4 Bounded evidence-gap allocation

For a rollout of valid response length T_g, define the detached evidence-gap priority and its within-rollout mean:

$$u_{gt} = v_{gt}\,\mathrm{sg}[J_{gt}], \qquad \bar{u}_g = \frac{1}{T_g}\sum_{t=1}^{T_g} u_{gt}, \qquad \alpha_{g,:} = \mathrm{sg}[\mathcal{B}(1+\rho(u_{g,:}-\bar{u}_g))]\,. \tag{8}$$

The operator $\mathcal{B}$ normalizes the mean to one and, if necessary, contracts deviations from one until all entries lie in $[0.5, 2]$; its exact implementation is given in Appendix A. A token receives more weight when the teacher depends on the visual input and the student still disagrees with the positive teacher. Uncertainty or disagreement alone is insufficient to raise this priority.

For rollout allocation, the score is the student gap averaged over teacher visual dependence, rather than a sum that grows automatically with response length:

$$R_g = \begin{cases} \dfrac{\sum_t v_{gt}\,\mathrm{sg}[J_{gt}]}{\sum_t v_{gt}}, & \sum_t v_{gt} > \epsilon, \\ 0, & \text{otherwise}, \end{cases} \qquad \beta_{i,:} = \mathrm{sg}\big[\mathcal{B}\big(1+\rho(R_{i,:}-\bar{R}_i)\big)\big]\,, \tag{9}$$

where $\bar{R}_i = G^{-1}\sum_g R_{ig}$ is computed among the G trajectories for the same question. We use the same $\rho = 0.5$ for both allocation levels. Thus, $T_g^{-1}\sum_t \alpha_{gt} = 1$ and $G^{-1}\sum_g \beta_{ig} = 1$. This creates a fixed per-rollout mass before importance correction, while allowing bounded redistribution within and across trajectories. It removes the direct token-count multiplier; it does not make the objective invariant to arbitrary additions of nonvisual text.

4.5 Training objective and procedure

Let c_{igt} denote detached importance corrections supplied by the training framework. The complete objective is

$$\mathcal{L} = \frac{1}{BG}\sum_{i=1}^{B}\sum_{g=1}^{G}\frac{\beta_{ig}}{T_{ig}}\sum_{t=1}^{T_{ig}}\alpha_{igt}c_{igt}\left[J_{igt} + \lambda v_{igt}H_{igt}\right]. \tag{10}$$

The exact coefficients of J_{igt} and each r_{igtk} are therefore

$$a_{igt}^{\mathrm{OPD}} = \frac{\beta_{ig}\alpha_{igt}c_{igt}}{BGT_{ig}}, \qquad a_{igtk}^{\mathrm{hidden}} = a_{igt}^{\mathrm{OPD}}\lambda v_{igt}\omega_{igtk}. \tag{11}$$

Padding contributes zero. Gradients flow only through student distributions and hidden states; teacher outputs, gates, priorities, allocation weights, scales, and importance factors are detached. The sampled trajectories are treated as fixed replay data during this update, with no extra policy-gradient reward term. Appendix C details the importance factors and distributed reduction.

Algorithm 1 Visual causal transport distillation: one actor update

Require: Student θ, EMA teacher ϕ, B image–question pairs and privileged views
Ensure: Updated student θ and teacher ϕ

$\mathcal{Y} \leftarrow \{y_i^{(g)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i, q_i)\}_{i=1,g=1}^{B,G}$ ▷ full-image rollouts
$(p^S, h^S) \leftarrow \text{Replay}_\theta(x, q, \mathcal{Y})$ ▷ student gradients enabled
$(p^+, h^+) \leftarrow \text{sg}[\text{Replay}_\phi(C^+, \mathcal{Y})]$
$(p^-, h^-) \leftarrow \text{sg}[\text{Replay}_\phi(C^+, \mathcal{Y}; \text{do}(E_{\text{vis}} = 0))]$
$J, v \leftarrow \text{SharedSupportJSD}(p^S, p^+, p^-)$ ▷ Eq. (2); v detached
for each valid response position (i, g, t) **do**
  Form (A_k^S, A_k^+, A_k^-) and ω_k for $k = 1, 2, 3$ ▷ Eqs. (3)–(5)
  $d_k^\pm \leftarrow \text{RMS}(A_k^S - A_k^\pm)$; $H_{igt} \leftarrow \sum_k \omega_k b_k \dfrac{d_k^+}{d_k^+ + d_k^- + \epsilon}$
end for
$(\alpha, \beta) \leftarrow \text{sg}[\text{EvidenceGapAllocation}(v, J)]$ ▷ Eqs. (8), (9)
$\mathcal{L} \leftarrow \frac{1}{BG} \sum_{i,g} \frac{\beta_{ig}}{T_{ig}} \sum_t \alpha_{igt} c_{igt} (J_{igt} + \lambda v_{igt} H_{igt})$
$\theta \leftarrow \text{OptimizerStep}(\theta, \nabla_\theta \mathcal{L})$; $\phi \leftarrow (1 - \tau)\phi + \tau\theta$

5 Understanding the Ratio Objective

The ratio has an interpretable behavior between the two teacher references. Fix a valid object and let $a \in [0, 1]$ parameterize the segment from the negative to the positive teacher: $A^S = A^- + a(A^+ - A^-)$. Write $D = \text{RMS}(A^+ - A^-) > 0$.

Proposition 1 (Relative progress on the teacher segment)**.** *Under the conditions above,* $r = (1 - a)D/(D + \epsilon)$. *When* $D \gg \epsilon$, *the penalty is approximately the remaining fraction* $1 - a$ *of the segment to the positive teacher.*

The proof follows by substituting $d^+ = (1 - a)D$ and $d^- = aD$; Appendix B supplies further properties. This statement describes the objective's geometry, not the actual optimization trajectory. In contrast, positive squared distance along this segment is $(1 - a)^2 D^2$, whose scale changes with reference separation. The ratio is consequently a natural measure of *relative* alignment, although this does not by itself establish an empirical advantage over MSE.

For fixed teacher references, $0 \leq H_t < 1$. If all valid objects are unchanged by the intervention, then $\omega_k = 0$ and $H_t = 0$. If $v_t = 0$, the hidden contribution in Eq. (10) is also zero. The two safeguards are complementary: ω determines which hidden objects carry intervention effects, while v determines whether the teacher's output prediction changes. A bounded loss does not guarantee bounded gradients; small reference distances can still produce large derivatives. The implementation retains standard gradient clipping and logs the constituent distances.

6 Experiments

Setup and evaluation. We evaluate Qwen3.5-4B and Qwen3.5-9B, comparing the initial checkpoint, OPSD, DAPO, Vision-OPD, and VCT. Following Vision-OPD (Yuan et al., 2026), the student is trained on full images from Vision-OPD-6K, while the teacher receives the corresponding privileged region views. All models are trained for one epoch on 8 NVIDIA H20 GPUs with 141 GB memory each, using a global batch of 96 prompts, eight rollouts per prompt, a learning rate of 2×10^{-6}, and ten warmup steps. Further implementation details are provided in Appendix D.

We evaluate on HR-Bench 4K/8K (Wang et al., 2024), MMStar (Chen et al., 2024a), TreeBench (Wang et al., 2026), Zoom (Wei et al., 2026), and MathVista (Lu et al., 2024), containing 800, 800, 1,500, 405, 845, and 460 examples, respectively. We report accuracy (%) on each benchmark and their unweighted average.

Main results. As shown in Table 1, VCT achieves the highest average accuracy at both model scales. On Qwen3.5-4B, it reaches 72.10%, improving over Vision-OPD by 1.39 points and outperforming

Table 1: Accuracy (%) on six multimodal benchmarks. Avg. denotes the unweighted mean across all benchmarks. Bold indicates the best result within each backbone, including ties.

Method	HR-4K	HR-8K	MMStar	TreeBench	Zoom	MathVista	Avg.
Qwen3.5-4B							
Initial	81.25	78.50	74.87	48.40	51.12	75.43	68.26
OPSD	81.75	77.00	74.33	48.64	61.07	68.26	68.51
DAPO	81.50	78.12	72.40	49.14	55.38	71.30	67.97
Vision-OPD	83.25	79.50	76.87	48.89	61.42	74.35	70.71
VCT (ours)	**84.50**	**83.00**	**78.33**	**49.38**	**61.54**	**75.87**	**72.10**
Qwen3.5-9B							
Initial	87.00	81.75	75.40	50.62	53.25	78.70	71.12
OPSD	83.50	81.50	75.53	51.36	63.08	75.22	71.70
DAPO	87.38	82.25	78.87	**51.60**	57.87	79.57	72.92
Vision-OPD	86.25	83.88	75.53	48.64	64.26	70.65	71.54
VCT (ours)	**87.75**	**86.50**	**80.40**	**51.60**	**64.62**	**79.78**	**75.11**

Table 2: Qwen3.5-4B ablations on HR-Bench 4K/8K and MMStar. Accuracy and Avg. are percentages; Avg. is the equal-weight mean of these three benchmarks, distinct from the six-benchmark average in Table 1. Results use the reported percentages. The no-hidden-loss run is pending. Bold marks the best reported value in each column.

Variant	λ	ρ	HR-4K	HR-8K	MMStar	Avg.
Default	1.0	0.5	84.50	**83.00**	**78.33**	**81.94**
No hidden loss	0.0	0.5	84.25	82.38	77.07	81.23
Uniform allocation	1.0	0.0	84.50	82.12	72.73	79.78
Lower allocation strength	1.0	0.25	85.50	82.50	77.40	81.80
Larger hidden coefficient	2.0	0.5	84.50	82.88	77.07	81.48
No rollout weight	1.0	0.5	84.38	81.62	76.67	80.89
No token weight	1.0	0.5	**86.12**	82.00	77.27	81.80

it on all six benchmarks. On Qwen3.5-9B, VCT achieves 75.11%, exceeding Vision-OPD by 3.57 points and DAPO by 2.19 points on average. The largest gains over Vision-OPD appear on MathVista (+9.13) and MMStar (+4.87). Overall, VCT leads all six benchmarks at 4B and five at 9B, while tying DAPO on TreeBench.

6.1 Ablation Study

Table 2 studies the effect of the hidden-loss coefficient λ and the allocation strength ρ on Qwen3.5-4B. The default setting is $(\lambda, \rho) = (1, 0.5)$.

Coefficient sensitivity. The default configuration achieves the best average accuracy of 81.94%. Reducing ρ to 0.25 gives 81.80%, while uniform allocation ($\rho = 0$) decreases the average to 79.78%. Increasing λ to 2 yields 81.48%, indicating that stronger hidden alignment does not further improve the aggregate performance.

Token and rollout allocation. Removing rollout-level weighting reduces the average to 80.89%, while removing token-level weighting gives 81.80%. Although the latter performs best on HR-4K, the full configuration achieves the strongest average across the evaluated benchmarks, supporting the use of joint token- and rollout-level evidence-gap allocation.

6.2 Qualitative analysis

Figure 4 presents a qualitative comparison between VCT and Vision-OPD on the same image and spatial-relation question. Both models identify the relevant visual entities, namely the pink banner and the green trash bin, indicating that the failure is not simply due to missing the queried objects. However, Vision-OPD reverses their horizontal relation and predicts that the pink banner is to the

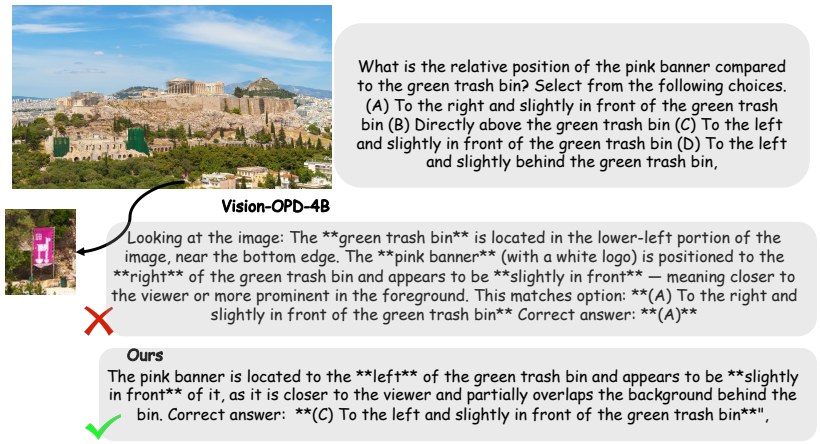

Figure 4: Qualitative comparison between Vision-OPD and VCT on a spatial-relation question. Both models identify the queried pink banner and green trash bin, but Vision-OPD reverses their horizontal relation and predicts that the banner is to the right of the bin. In contrast, VCT correctly infers that the pink banner is to the left and slightly in front of the green trash bin, matching the reference answer.

right of the green trash bin, leading to an incorrect answer. In contrast, VCT correctly determines that the pink banner is to the left and slightly in front of the green trash bin, consistent with the reference answer.

This example illustrates the distinction between recognizing visually associated concepts and using visual evidence to resolve their relative configuration. The baseline can localize the salient objects but does not preserve the correct spatial relation between them, whereas VCT produces a response that is more consistent with the image-conditioned relation required by the question. This behavior is consistent with the motivation of our intervention-based supervision: rather than transferring only the teacher's output preference, VCT encourages the student to follow representation changes induced by visual evidence and to retain such evidence-dependent information during reasoning.

7 Limitations and Broader Impact

Feature zeroing may be out of distribution, and the sampled prefix can already contain image-derived information. The teacher contrast therefore measures residual visual dependence conditional on that prefix. Imperfect crops and teacher errors can affect supervision. Training adds one negative teacher language-model pass and selected hidden-state storage.

8 Conclusion

We presented VCT, an on-policy self-distillation method that makes the teacher's visual intervention response part of hidden supervision. A single intervened teacher provides a counterfactual reference for relative alignment of final representations and cross-depth updates. Intervention energies and bounded evidence-gap weights determine where this supervision is applied. Across six benchmarks and two model sizes, VCT improves average accuracy over Vision-OPD by 1.39 and 3.57 percentage points without changing student inference.

AI USE STATEMENT

Generative AI assistance was used to inspect the existing implementation, translate it into mathematical notation, formulate and explain the objective's geometric properties and proofs, identify relevant literature, revise the 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. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? arXiv:2403.20330, 2024a. URL https://arxiv.org/abs/2403.20330.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5008–5017, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Distilling_Knowledge_via_Knowledge_Review_CVPR_2021_paper.html.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2011.10566.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6621–6642, 2024b. URL https://proceedings.mlr.press/v235/chen24j.html.

Zebang Cheng, Shuimu Chen, Boxue Yang, Yuanshen Guan, Jingyi Chen, Zheng Lian, Xiaojiang Peng, Fei Ma, LaiZhong Cui, and Qi Tian. OmniOPSD: Rationale-privileged on-policy self-distillation for affective computing. arXiv:2606.15920v2, 2026. URL https://arxiv.org/abs/2606.15920v2.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616, 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.07733.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.08543.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1911.05722.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. arXiv:2601.20802, 2026. URL https://arxiv.org/abs/2601.20802.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.07079.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

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 *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31044–31062, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.16922.

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. arXiv:2602.02994, 2026a. URL https://arxiv.org/abs/2602.02994.

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:2608.14144, 2026b. URL https://arxiv.org/abs/2608.14144.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-policy self-distillation without any supervision. arXiv:2608.06296, 2026c. URL https://arxiv.org/abs/2608.06296.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in hilbert space. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2002.05715.

Richard Yuanzhe Pang and He He. Text generation by learning from demonstrations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.07839.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1904.05068.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. arXiv:2209.15189, 2022. URL https://arxiv.org/abs/2209.15189.

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. arXiv:2606.25319, 2026. URL https://arxiv.org/abs/2606.25319.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1703.01780.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.10699.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, volume 2026, pp. 148769–148794, 2026.

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:2408.15556, 2024. URL https://arxiv.org/abs/2408.15556.

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. arXiv:2602.11858, 2026. URL https://arxiv.org/abs/2602.11858.

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:2605.18740, 2026. URL https://arxiv.org/abs/2605.18740.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. arXiv:2401.10020, 2024. URL https://arxiv.org/abs/2401.10020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.03928.

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:2607.28590, 2026. URL https://arxiv.org/abs/2607.28590.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Zhang_Be_Your_Own_Teacher_Improve_the_Performance_of_Convolutional_Neural_ICCV_2019_paper.html.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2203.08679.

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:2601.18734, 2026. URL https://arxiv.org/abs/2601.18734.

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:2607.24447, 2026. URL https://arxiv.org/abs/2607.24447.

A Exact Allocation and Probability Computation

Bounded mean-one normalization. Consider positive weights w_j over a nonempty valid set $\mathcal{V}$, with lower and upper bounds $0 < l \le 1 \le u$. The function `bounded_mean_one` first computes

$$\mu = |\mathcal{V}|^{-1} \sum_{j \in \mathcal{V}} w_j, \quad z_j = w_j / \max(\mu, 10^{-12}), \quad \delta_j = z_j - 1,$$
$$a = \max_{j \in \mathcal{V}} \max(\delta_j, 0), \qquad b = \max_{j \in \mathcal{V}} \max(-\delta_j, 0),$$
$$\eta = \min\left(1, \frac{u-1}{\max(a, 10^{-12})}, \frac{1-l}{\max(b, 10^{-12})}\right), \qquad \mathcal{B}(w)_j = 1 + \eta\delta_j. \tag{12}$$

Padding entries are zero. With the default raw weights, μ is positive and far above the numerical floor. Since $\sum_j \delta_j = 0$, valid weights retain mean one. Multiplying every deviation by a common η enforces the bounds without the bound violations that can arise from clipping followed by renormalization. All quantities are detached.

In the current evidence-gap configuration, $u_{gt}, R_g \in [0, 1]$. With $\rho = 0.5$, centered raw weights already lie in $[0.5, 1.5]$ and have mean one, up to roundoff. The wider upper bound of two is a safeguard, not evidence that the implementation regularly gives individual tokens a weight of two. Changing bounds or ρ changes this behavior and must be recorded.

Shared vocabulary partition. Let $\mathcal{K}_t$ be the K highest-probability entries under the current student at a response prediction position. For each branch b, construct

$$\tilde{p}_t^b = \left((p_t^b(v))_{v \in \mathcal{K}_t},\ 1 - \sum_{v \in \mathcal{K}_t} p_t^b(v) \right). \tag{13}$$

Teacher probabilities are gathered at *student-selected* indices. The tail is included for all branches, retaining the omitted mass as one category. This is not teacher-top-K distillation or renormalization of just the selected entries. The coarsening can conceal disagreements within the tail, so v_t is a support-dependent proxy. The top-K selection is discrete; gradients flow through selected student probabilities and the tail, not through the selection of indices.

For numerical stability, the code computes selected log mass by log-sum-exp, caps it above at -10^{-7}, and obtains log tail mass as $\log(-\operatorname{expm1}(\log m))$. The JSD routine renormalizes each coarsened log distribution with log-sum-exp, computes in float32, divides by $\log 2$, and clamps the result to $[0, 1]$. The cap can make negligible differences from exact probability arithmetic. Distribution logits use the configured sampling temperature, which is one in the inherited defaults. The $\epsilon = 10^{-6}$ used in hidden objectives and allocation is distinct from this tail-mass numerical safeguard.

B Properties of Counterfactual Relative Alignment

Assume finite aligned hidden tensors, $\epsilon > 0$, and a fixed teacher during a student update. All statements below concern the objective as written, not a guarantee about benchmark performance.

Proof of Proposition 1. For $A^S = A^- + a\Delta$ with $\Delta = A^+ - A^-$ and $a \in [0, 1]$, homogeneity of the Euclidean norm gives $d^+ = (1-a)\,\mathrm{RMS}(\Delta)$ and $d^- = a\,\mathrm{RMS}(\Delta)$. Substitution into Eq. (7) gives $r = (1-a)D/(D+\epsilon)$. For $D \gg \epsilon$, the difference from $1-a$ is $(1-a)\epsilon/(D+\epsilon)$. The result assumes a valid object, so $b_k = 1$. □

Lemma 1 (Boundedness and zero-intervention behavior)**.** *For the implemented objective,* $0 \le r_k < 1$, $0 \le \sum_k \omega_k < 1$, *and* $0 \le H_t < 1$. *If* $q_k = 0$ *for every object, then* $H_t = 0$. *If the student matches all positive-teacher objects, then* $H_t = 0$.

Proof. All distances and energies are nonnegative. The denominator of r_k is strictly larger than its numerator when the numerator is positive; if it is zero, $r_k = 0$. Similarly, writing $Q = \sum_k q_k$, the sum of object weights is $Q/(Q+\epsilon) < 1$. A nonnegative weighted sum of values below one is below one. When $Q = 0$, each object weight is zero. At the positive teacher, each positive distance is zero. Invalid objects are explicitly masked and do not alter these conclusions. □

Gradient interpretation. For a valid object, treating its distances as scalar variables,

$$\frac{\partial r}{\partial d^+} = \frac{d^- + \epsilon}{(d^+ + d^- + \epsilon)^2}, \qquad \frac{\partial r}{\partial d^-} = -\frac{d^+}{(d^+ + d^- + \epsilon)^2}. \tag{14}$$

Both terms contribute to the student gradient. Away from zero distances,

$$\nabla_{A^S} r = \frac{(d^- + \epsilon)\nabla d^+ - d^+ \nabla d^-}{(d^+ + d^- + \epsilon)^2}, \qquad \nabla d^{\pm} = \frac{A^S - A^{\pm}}{\sqrt{d}\,\|A^S - A^{\pm}\|_2}. \tag{15}$$

The implementation uses the vector norm divided by $\sqrt{d}$, with its finite zero subgradient at equality, rather than a naive square root of an MSE that can give undefined backward behavior at zero. The geometry is attraction toward the positive reference relative to distance from the negative reference. For very distant student states, the ratio can approach one half and provide weak progress information; the output loss remains important. For nearly coincident teacher references, derivatives can be large even though the value is bounded. Energy weighting and ϵ mitigate degenerate cases but do not supply a universal gradient-norm bound.

Common-unit invariance. If all three raw branches of one object are multiplied by the same positive constant and its validity does not change, the teacher RMS scale multiplies by that constant and every A_k^b is unchanged. The normalized ratio and intervention energy are then unchanged. This property comes from the common teacher scale; a teacher-scaled MSE shares it. The additional property of the ratio is its reference-relative geometry, not scale invariance unique to this paper.

What is and is not causal transport. The negative teacher implements a controlled intervention on projected image embeddings while preserving the model, textual prefix, and token layout. For the depth-update object, the response is a difference of intervention effects across layers, as Eq. (6) shows. There is no assumption that the prefix is independent of the image: it was generated from a full-image student. Consequently, the contrast estimates a model response conditional on that prefix. It cannot establish a natural direct effect of image pixels, distinguish all visual from textual information, or prove generalization under arbitrary image shifts. Since the student has no intervened branch, the loss also does not enforce $\Delta^S = \Delta^T$. It transfers teacher representations and updates *using* a teacher intervention as a reference.

C Implementation Fidelity and Reduction

Layer capture. The selected outputs are captured after blocks $\lceil 0.25L \rceil$, $\lceil 0.5L \rceil$, and L. In zero-based module indexing, these are indices $\lceil fL \rceil - 1$. The final selected state is pre-final-normalization, so it is not the post-normalization tensor immediately consumed by the language-model head. Hidden objects are formed in float32. Teacher tensors, normalization scales, and component weights are detached. The three branches have the same hidden dimension; no learned projector is used.

Negative input path. In the Qwen3.5 model adapter, ordinary image processing produces projected visual features that are inserted into image-token slots. The negative branch inserts zeros of the same shape at those slots and skips the visual encoder. It retains the positive teacher's image grid metadata, positions, masks, and text. The checked implementation supports image inputs for this intervention and validates the visual slot count. Describing this operation as black-box pixel masking, a black crop, or deletion of the image-token sequence would not match the code.

Importance factors. When supplied, rollout correction weights account for differences between the actor's recomputed probabilities and those recorded by the rollout engine. In the default token-level setting the factor is a detached exponentiated log-probability difference, upper-clipped at two. The actor can additionally multiply by

$$c_t^{\text{actor}} = \text{sg}[\min(2, \exp\{\text{clip}(\log \pi_\theta(y_t) - \log \pi_{\text{old}}(y_t), -20, 20)\})]\,. \tag{16}$$

For the single-minibatch, single-epoch on-policy update path, the actor uses its current detached log probability as the old value, making this additional factor one. The effective c is the product of whichever factors are active; if none is supplied, it is one. These are implementation corrections, not additional visual-allocation coefficients. The bounds on α and β do not bound their product with arbitrary importance factors by the same interval. Exact engine correction settings must be retained in the experiment configuration.

Distributed objective. For D data-parallel ranks, the schedule constructs complete rollout groups in waves of $\mathrm{lcm}(G, D)$ trajectories. Each rank processes $G/\gcd(G, D)$ trajectories per wave. Only detached rollout scores and identifiers are gathered for β; this allocation step does not gather hidden tensors. With $G = D = 8$, a synchronized wave has one trajectory on each rank. This is a microbatching decision, not a global batch-size change: a 96-prompt batch still contains 768 sampled trajectories.

The local loss is a mean over microbatch sequences. Before backward, it is scaled by the local microbatch row count divided by the local minibatch row count. Data-parallel gradient averaging recovers the global sequence-mean objective in Eq. (10) when the scheduled rank workloads are balanced. Each sequence already contains its $1/T_g$ factor; there must not be a second division by length. Debug coefficient tensors use the global trajectory count explicitly, as Eq. (11) does.

D Configuration and Reproducibility Record

Table 3: Configuration defaults.

Setting	Default
Hidden objective / allocation	`causal_ratio` / `evidence_gap`
New loss balance λ / allocation strength ρ	1.0 / 0.5
Hidden capture fractions	$[0.25, 0.5, 1.0]$
Token / rollout bounds	Both $[0.5, 2.0]$
Hidden and allocation ϵ	10^{-6}
Teacher update	EMA; new student contribution 0.05
Output divergence	Symmetric JSD divided by $\log 2$
Vocabulary approximation	Student top-100 + one tail bucket
Backbone selected by launcher	Qwen3.5-9B; model path overridable
Global prompt batch / rollouts per prompt	96 / 8
Actor prompt minibatch / actor epochs	96 / 1
Learning rate / warmup	2×10^{-6} / 10 steps
Training duration	One epoch
Maximum student prompt / response length	8192 / 1024 tokens
Default node / GPU count	1 / 8
Checkpoint frequency / resume	Every 10 steps / disabled
Actor aggregation in this branch	Sequence mean of token means
Actor entropy calculation	Disabled in launcher
Extra reward / actor KL objective	Disabled
Dynamic batching / gradient checkpointing	Enabled / enabled
FSDP parameter / optimizer offload	Enabled / enabled
Detailed debug frequency	Every step, first permitted microbatch
Detailed debug selection	Rank zero; up to 8 rollouts, 16 tokens