SEE WHAT I SEE: ON-POLICY DISTILLATION OF VISUAL EVIDENCE, NOT JUST WORDS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has become a core technique in the post-training of large language models and is increasingly applied to multimodal ones. It trains a student with dense token-level feedback from a teacher on the student's own rollouts. However, this feedback supervises what the student says, not what it sees: even a correct token may draw little from the image, or draw from the wrong region. We find that during OPD the student's output distribution converges to the teacher's, while an evidence gap persists. The gap has two sides: (i) a few load-bearing tokens carry most of the teacher's evidence, yet their supervision is diluted across the whole response; (ii) on these tokens the student's evidence misses the ground-truth region, even where its next-token distribution already matches the teacher's. To close this gap, we propose SWIS (*See What I See*), which distills the teacher's visual evidence directly rather than through its next-token distribution alone. SWIS models evidence as a joint distribution over token–patch pairs and aligns the student's with the teacher's, matching both which tokens carry the evidence and where on the image it lands. Across ten benchmarks, SWIS improves average accuracy by 1.8 pp, gaining most where answers depend on the image, and closes the evidence gap under the same readout that exposed it.

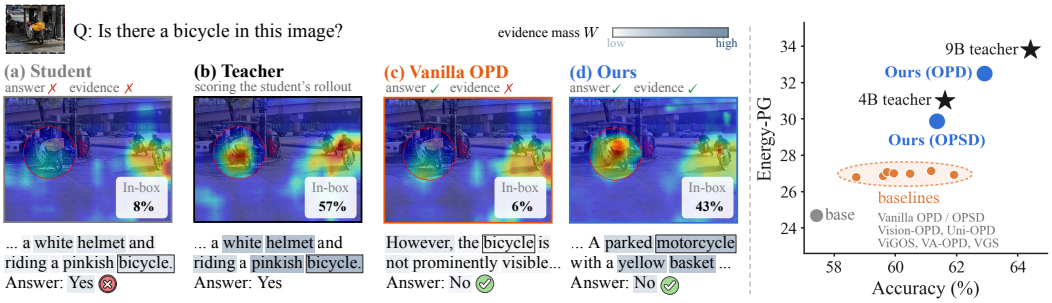


Figure 1: **OPD transfers what the teacher says, not what it sees.** *Left:* The teacher looks at the annotated region (b). However, the student does not, even after vanilla OPD (a, c); SWIS makes it look (d). Heatmaps show the evidence of the boxed word. *Right:* Baselines raise held-out accuracy, but their Energy-PG (response-level In-box) stays well below the teachers'; SWIS raises both.

1 INTRODUCTION

On-policy distillation (OPD) has emerged as a promising post-training paradigm that supervises a student on its own rollouts with dense token-level feedback from a teacher, avoiding the exposure bias of supervised fine-tuning and the sparse outcome rewards of reinforcement learning (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025; Li et al., 2026c). With the rise of multimodal large language models (MLLMs), OPD has been extended to vision-language students, either with a stronger external teacher (Hou et al., 2026; Bousselham et al., 2026; Li et al., 2026b) or with a privileged self-teacher that conditions the same model on a reference solution, an evidence crop, or a textual visual cue (Zhao et al., 2026; Yuan et al., 2026; Tian et al., 2026). Recent work has further demonstrated the effectiveness of evidence-aware objectives that reweight, gate, or reshape

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.

the output loss by each token's visual dependency (Liu et al., 2026; Yoon et al., 2026; Liang et al., 2026; Sun et al., 2026; Qian et al., 2026; Zhang et al., 2026).

Despite these advances, existing multimodal OPD methods fundamentally refine the student's next-token distribution toward the teacher's, under the implicit assumption that a student who matches it also inherits how the teacher uses the image. We argue that this assumption fails: next-token matching supervises only the result (the token), not the process (the visual evidence behind it). A token earns the same credit whether it was read off the relevant region, a background shortcut, or a language prior (McCoy et al., 2019; Hao et al., 2025; Parcalabescu & Frank, 2025; Guan et al., 2024), so the objective constrains neither how much evidence each token draws nor where on the image it lands. Supervising the visual side directly does not supervise this evidence either: a perception reward (Wang et al., 2026b; Xiao et al., 2026; Li et al., 2026d; Gong et al., 2026) scores the whole response rather than any single token, and attention alignment (Li et al., 2026a; Kim et al., 2026) matches maps normalized over positions, which record where a token looked but not how much it depended on the image. This raises the central concern of this paper:

In on-policy distillation, does matching what the teacher says transfer what it sees?

Through empirical analysis, as illustrated in Fig. 1, we find that the answer is no: as training progresses, the student's next-token distribution converges to the teacher's and its accuracy rises, yet the share of its visual evidence on the annotated region does not follow. We call this the *evidence gap*: the student's visual evidence remains far from the teacher's even after their outputs converge. Next-token matching constrains no property of the evidence; the student gains accuracy where shortcuts or priors already suffice, but not where the answer depends on the image. On the token side, a small minority of tokens carries most of the teacher's evidence, and these load-bearing tokens are the ones the student under-weights and next-token matching under-supervises. This raises our first question: **I) Mass:** ***How can we supervise how much evidence each token draws from the image?*** On the image side, the student's evidence on these tokens lands off the annotated region and stays there, even where its next-token distribution already matches the teacher's. This raises our second question: **II) Placement:** ***How can we supervise where on the image that evidence is drawn from?***

To address these questions, we propose **S**ee **W**hat **I S**ee (**SWIS**), which matches the student's evidence to the teacher's, token by token, alongside the next-token objective. We measure evidence as the contribution of each image patch to each token logit (Kobayashi et al., 2020), computed in the same forward pass, and normalize it over the response into a joint distribution over token–patch pairs. For I), we constrain its marginal, the evidence mass each token draws; for II), its conditional, where that mass lands on the image. By the chain rule, the two constraints are exactly the mass and placement terms of a single KL divergence on this distribution, so one divergence answers both questions, with no coefficient between mass and placement and no extra forward or backward pass. By extending the distillation target from the next-token distribution to the evidence behind it, SWIS transfers what the teacher sees along with what it says. Our main contributions are as follows:

❶ We identify a persistent problem in multimodal OPD, termed the *evidence gap*: next-token matching drives the student's predictions and accuracy toward the teacher's, but its visual evidence stays far from the teacher's, so its accuracy gains largely bypass the region that holds the answer.

❷ We propose SWIS, which extends the distillation objective from next-token matching to evidence matching, enabling direct supervision of visual grounding.

❸ Across ten benchmarks, two student scales, and two teacher settings, SWIS achieves the best average accuracy in all four settings, **1.8** points above its vanilla recipe on average. Over vanilla OPD, it also raises the share of its evidence on the annotated region by **5.4** points.

2 Related Work

Knowledge distillation trains a compact student to match a stronger teacher's output distribution (Hinton et al., 2015). On-policy distillation moves this supervision onto the student's own samples so that per-token corrections land on states the student actually visits (Gu et al., 2024; Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025; Li et al., 2026c), and has been adopted in production post-training pipelines (Qwen Team, 2026a) and extended to multimodal students (Hou et al., 2026; Bousselham et al., 2026; Li et al., 2026b). When no external teacher is available, on-policy self-distillation (OPSD) builds one from the student itself by conditioning on train-time privileges such

as reasoning traces (Zhao et al., 2026), regional crops (Yuan et al., 2026), or evidence descriptions (Tian et al., 2026). In all cases the supervision targets only the output distribution, token by token.

Several lines of work push supervision toward the visual side: dependency weighting adjusts how much each token's loss should rely on the image (Huang et al., 2026; Liu et al., 2026; Ye et al., 2026; Yoon et al., 2026); attention alignment supervises where to attend (Li et al., 2026a; Kim et al., 2026); RL methods reward perception quality directly (Wang et al., 2026b; Gong et al., 2026). Each addresses one axis; none jointly constrains which patches support each token. We distill a per-token evidence distribution over image patches, supervising all three at once.

3 Methodology

3.1 Preliminaries

On-policy distillation. Consider a multimodal task with an image $\boldsymbol{I}$ and a question $\boldsymbol{q}$. OPD trains a student policy π_{S} to match a teacher π_{T} on the student's own generations: sampling a response $\boldsymbol{y} = (y_1, \ldots, y_{|\boldsymbol{y}|}) \sim \pi_{\mathrm{S}}(\cdot \mid \boldsymbol{I}, \boldsymbol{q})$, it minimizes a sequence-averaged reverse KL between the two policies' next-token distributions,

$$\mathcal{L}_{\mathrm{out}} = \mathbb{E}_{(\boldsymbol{I},\boldsymbol{q})\sim\mathcal{D},\, \boldsymbol{y}\sim\pi_{\mathrm{S}}(\cdot\mid\boldsymbol{I},\boldsymbol{q})}\Big[\tfrac{1}{|\boldsymbol{y}|}\textstyle\sum_{t=1}^{|\boldsymbol{y}|} D_{\mathrm{KL}}\big(\pi_{\mathrm{S}}(\cdot \mid \boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y}_{<t}) \,\big\|\, \pi_{\mathrm{T}}(\cdot \mid \boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y}_{<t})\big)\Big]. \tag{1}$$

OPD combines the on-policy relevance of reinforcement learning with the dense token-level guidance of distillation, making it a natural fit for autoregressive generation and reasoning.

Reading out visual evidence. To quantify where in the image each token draws its support, we define an evidence score $e(t, j)$ that decomposes each token's logit into per-patch contributions, accumulated over layers ℓ and heads h:

$$e(t,j) = \Big[\textstyle\sum_{\ell,h} \alpha_{t,j}^{(\ell,h)} \big\langle f^{(\ell,h)}(\boldsymbol{z}_j^{(\ell-1)}),\, \boldsymbol{u}_{y_t}\big\rangle\Big]_+, \qquad f^{(\ell,h)}(\boldsymbol{z}) = \boldsymbol{z}\,\boldsymbol{W}_{\mathrm{V}}^{(\ell,h)}\boldsymbol{W}_{\mathrm{O}}^{(\ell,h)}, \tag{2}$$

where $\alpha_{t,j}^{(\ell,h)}$ is the attention weight from the position predicting y_t to patch j, $f^{(\ell,h)}$ is the head's value pathway applied to the patch representation $\boldsymbol{z}_j^{(\ell-1)}$, and $\boldsymbol{u}_{y_t}$ is the unembedding direction of y_t. Each term multiplies how strongly the model attends to a patch by whether what it reads supports the token, so e measures evidential use of the image rather than attention (Kobayashi et al., 2020).

3.2 A Closer Look at On-Policy Distillation

To test whether matching next-token distributions also transfers the visual evidence behind them, we train a student with Eq. (1) alone under both OPD and OPSD and track task accuracy, next-token divergence from the teacher, and where the student's evidence falls on the image. We find a persistent *evidence gap* in both settings: next-token matching improves the answers but does not move the evidence. We examine the gap first over training, whether the evidence follows the outputs as they converge (Observation 1), and then from the two sides of the evidence, which tokens carry it and where on the image it lands (Observations 2 and 3).

Observation 1: Improved answers, persistent evidence gap. As shown in Fig. 2, under both OPD and OPSD, the student's output divergence drops early and its accuracy approaches the teacher's. Energy-PG (Wang et al., 2020), the share of the student's evidence that lands inside the annotated region, does not follow. It stays near its starting value, only slightly above chance (the region's share of the image area) and far below the teacher's. The gap is not an artifact of the readout. Masking the annotated region leaves the student's newly correct answers largely intact, unlike the teacher's (Appendix C.2). Output matching therefore reproduces the teacher's predictions but not the evidence they rest on.

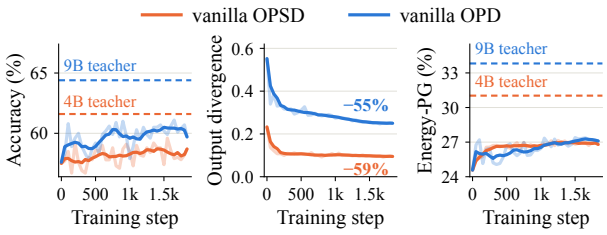
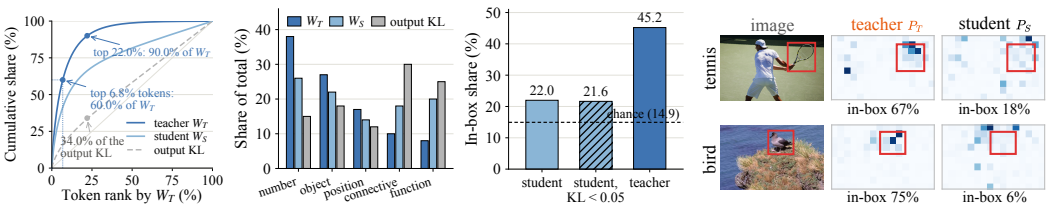


Figure 2: As accuracy rises and divergence falls, energy-PG stays far below the teachers'.



(a) Evidence focus (b) Token type (c) In-box share (d) Evidence maps

Figure 3: **The evidence gap has two sides.** The student under-weights the load-bearing tokens (a, b), and on those tokens its evidence lands off the region even where its next-token distribution already matches the teacher's (c, d). In (a) tokens are ranked by the teacher's share of evidence W_T, with the student's W_S and the share of the output loss along the same ranking; (b) groups the same shares by token type; (d) marks the annotated box in red.

To locate the gap, we further read Eq. (2) from two sides. Along the response we ask which tokens draw on the image, and how much; inside the image we ask where each token's evidence lands.

Observation 2: Sparse evidence, missed by the output loss. On the same student-generated response we measure each token's share of the total evidence, $W(t) = \sum_j e(t,j) / \sum_{t',j'} e(t',j')$, for both teacher and student. Evidence is sparse. A small minority of tokens carries most of the teacher's evidence (Fig. 3a), and these load-bearing tokens, the top few percent by W_T, are exactly where the student puts too little weight; in particular it under-weights the numbers and attributes that have to be read off the image (Fig. 3b). The output loss does not correct this allocation. Its per-token KL is only weakly concentrated on the load-bearing tokens (Fig. 3a), and most of it falls on reasoning connectives and function words. The tokens that need the image are under-weighted by the student and under-supervised by the objective.

Observation 3: Spatial mismatch despite matched token distributions. For a token t we take $P(\cdot \mid t)$, the row $e(t,\cdot)$ normalized over patches, and measure the share that falls inside the annotated region on the load-bearing tokens. The teacher's evidence lands inside the region far more often than chance; the student's barely exceeds it (Fig. 3c). The mismatch survives output convergence. On the tokens whose next-token distribution already matches the teacher's, the student's rate is no better. The student has matched the word without matching the look. Fig. 3d shows two such tokens, where the teacher concentrates on the box and the student spreads across the image. Matching the next-token distribution supplies no spatial supervision.

These observations show that next-token matching aligns the student's answers with the teacher's but not the evidence behind them:

> **Evidence Gap**: *the divergence between the student's and the teacher's visual evidence, across tokens and across the image, that next-token matching leaves open.*

The remaining question is whether a student can learn what the teacher sees, not only what it says.

3.3 See What I See

To close the evidence gap, SWIS extends the distillation target from the teacher's next-token distribution to the evidence behind it (as shown in Fig. 4). The gap has two sides: the student under-weights the load-bearing tokens (Observation 2), and on them its evidence lands away from the teacher's (Observation 3). SWIS therefore matches the teacher on how much evidence each token draws (*evidence mass*) and where on the image it lands (*evidence placement*).

3.3.1 Evidence Mass and Placement

Both are read from the evidence score in Eq. (2), normalized in two ways. For a response, let

$$W(t) = \frac{\sum_j e(t,j)}{\sum_{t'} \sum_{j'} e(t',j')}, \qquad P(j \mid t) = \frac{e(t,j)}{\sum_{j'} e(t,j')}. \tag{3}$$

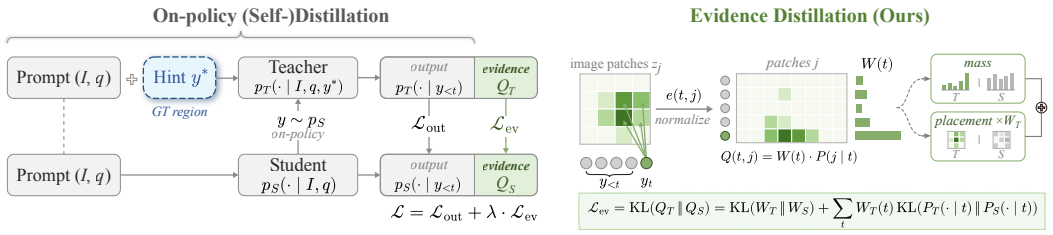


Figure 4: **On-policy (self-)distillation (left) and our evidence distillation (right).** Green parts are ours: SWIS keeps the next-token loss and, on the same forward passes, adds an evidence loss that matches the teacher on how much each token draws on the image and where.

Here the evidence mass $W(t)$ is the share of the response's evidence that token t draws, and the evidence placement $P(j \mid t)$ is the share of token t's evidence that comes from patch j. As in OPD, both are read on the student's own response, by the teacher (W_{T}, P_{T}) and by the student (W_{S}, P_{S}). Because W is normalized over the whole response rather than per token, a token that barely reads the image keeps near-zero mass, preserving the sparsity measured in Observation 2.

In order for the student to draw the same evidence as the teacher for every token, we constrain each quantity with a KL divergence and weight the placement term by the teacher's evidence mass,

$$\mathcal{L}_{\mathrm{mass}} = D_{\mathrm{KL}}(W_{\mathrm{T}} \,\|\, W_{\mathrm{S}})\,, \qquad \mathcal{L}_{\mathrm{place}} = \sum_t W_{\mathrm{T}}(t)\, D_{\mathrm{KL}}(P_{\mathrm{T}}(\cdot \mid t) \,\|\, P_{\mathrm{S}}(\cdot \mid t))\,. \tag{4}$$

The W_{T} weights follow from the same sparsity. A token that draws almost no evidence has a placement $P(\cdot \mid t)$ that is mostly noise, and matching such placements with equal weight would ask the student to reproduce noise. Weighting by W_{T} puts the constraint on the load-bearing tokens.

3.3.2 A Single Evidence Divergence

The W_{T} weights do more than suppress noise. They are exactly the weights under which the two constraints merge into a single divergence. Let Q be the evidence normalized as a whole, a joint distribution over tokens and patches,

$$Q(t, j) = W(t)\, P(j \mid t) = \frac{e(t, j)}{\sum_{t'} \sum_{j'} e(t', j')}, \tag{5}$$

the fraction of the response's evidence that token t draws from patch j. Write Q_{T} and Q_{S} for the teacher and the student.

Proposition 1 (Exact decomposition of the evidence divergence) *For Q_{T} and Q_{S} defined in Eq. (5),*

$$D_{\mathrm{KL}}(Q_{\mathrm{T}} \,\|\, Q_{\mathrm{S}}) = \underbrace{D_{\mathrm{KL}}(W_{\mathrm{T}} \,\|\, W_{\mathrm{S}})}_{\text{Mass term } \mathcal{L}_{\mathrm{mass}}} + \underbrace{\sum_t W_{\mathrm{T}}(t)\, D_{\mathrm{KL}}(P_{\mathrm{T}}(\cdot \mid t) \,\|\, P_{\mathrm{S}}(\cdot \mid t))}_{\text{Placement term } \mathcal{L}_{\mathrm{place}}}\,. \tag{6}$$

The identity is the chain rule of KL; a derivation is in Appendix B. Matching both sides of the gap thus reduces to one divergence on the joint evidence distribution, computed in one pass over the two evidence matrices. The identity also fixes the balance between mass and placement, so there is no coefficient to set between them.

SWIS uses this divergence as its evidence loss,

$$\mathcal{L}_{\mathrm{ev}} = D_{\mathrm{KL}}(Q_{\mathrm{T}} \,\|\, Q_{\mathrm{S}})\,. \tag{7}$$

We backpropagate only through Q_{S} and treat Q_{T} as a fixed target. During training, the student branch replaces the hard rectifier in Eq. (2) with Softplus, so Q_{S} stays strictly positive and Eq. (7) remains finite wherever the teacher places evidence. The objective adds this term to $\mathcal{L}_{\mathrm{out}}$,

$$\mathcal{L} = \mathcal{L}_{\mathrm{out}} + \lambda\, \mathcal{L}_{\mathrm{ev}}. \tag{8}$$

The teacher and the student each run one forward pass on the student-generated response. The logits give $\mathcal{L}_{\mathrm{out}}$; the tensors already used in Eq. (2) give Q_{T}, Q_{S}, and $\mathcal{L}_{\mathrm{ev}}$.

Table 1: **Comparison with state-of-the-art methods** on visual math, hallucination, and general visual understanding benchmarks, with Qwen3.5-4B and Qwen3.5-2B as base models. Best and second-best results are marked in **bold** and underline, respectively. * denotes our re-implementation.

Method	Visual Math			Hallucination		General Visual Understanding					Avg.
	MathVista	MathVerse	MathVision	HallusionBench	POPE	MMVP	MMMU-Pro	MMMU	MMStar	RealWorldQA	
Qwen3.5-4B											
Base Model	82.00	71.19	47.70	73.50	87.80	80.33	56.82	68.89	77.27	76.34	72.18
On-Policy Self-Distillation (OPSD)											
Vanilla OPSD	83.50	72.69	48.68	73.98	88.92	80.67	57.69	69.22	76.13	75.69	72.72
Vision-OPD	82.20	71.29	48.42	73.36	87.17	80.00	57.76	**70.22**	74.40	73.20	71.80
ViGOS	84.20	**72.94**	48.42	73.71	88.83	81.67	56.88	69.00	76.47	76.99	72.91
Ours	**84.90**	72.89	**48.91**	**75.91**	**90.92**	**83.33**	**57.92**	69.67	**78.07**	**77.25**	**73.98**
On-Policy Distillation (OPD)											
Teacher	88.30	79.62	55.95	78.29	91.78	85.00	64.50	74.89	81.93	80.52	78.08
Vanilla OPD	83.70	70.81	47.50	72.61	88.49	79.33	57.52	70.11	76.93	76.86	72.39
Uni-OPD	83.30	71.09	47.80	71.66	88.70	79.00	58.74	68.33	75.07	76.21	71.99
VA-OPD*	85.00	**72.21**	47.60	73.87	89.57	78.67	59.06	**71.67**	77.40	75.95	73.10
VGS*	84.10	71.70	**48.91**	72.29	89.52	76.00	59.31	69.56	78.13	76.47	72.60
Ours	**85.80**	71.90	48.59	**76.29**	**89.70**	**81.33**	**60.50**	71.22	**79.93**	**77.65**	**74.29**
Qwen3.5-2B											
Base Model	71.90	58.27	25.99	66.67	89.26	76.00	35.38	47.44	74.47	68.89	61.43
On-Policy Self-Distillation (OPSD)											
Vanilla OPSD	71.50	57.79	25.82	66.56	89.33	76.67	34.89	47.00	74.00	68.63	61.22
Vision-OPD	68.40	48.30	21.09	56.89	76.56	61.67	18.90	42.78	51.87	55.82	50.23
ViGOS	73.00	58.60	25.99	68.77	**89.40**	76.33	37.89	49.33	74.80	69.15	62.33
Ours	**73.20**	**61.29**	**29.01**	**69.72**	89.29	**79.00**	**44.40**	**50.22**	**75.87**	**69.54**	**64.15**
On-Policy Distillation (OPD)											
Teacher	88.30	79.62	55.95	78.29	91.78	85.00	64.50	74.89	81.93	80.52	78.08
Vanilla OPD	75.30	57.46	25.66	71.19	89.59	78.00	45.20	59.33	75.07	72.16	64.90
Uni-OPD	74.90	57.51	26.22	69.61	89.04	75.33	44.22	56.56	74.60	66.93	63.49
VA-OPD*	**76.70**	58.91	25.69	71.08	89.41	77.33	**45.38**	**61.00**	74.33	**73.07**	65.29
VGS*	75.60	58.40	**26.51**	68.24	89.18	77.00	41.68	59.22	73.80	72.03	64.17
Ours	76.30	**59.21**	26.41	**71.29**	**90.66**	**82.00**	44.90	60.22	**75.60**	72.94	**65.95**

3.3.3 Discussion

Standard OPD supervises only the next-token distribution. The two lines of work push this supervision toward the image, but each reaches only one side of the evidence. Dependency weighting (Huang et al., 2026; Liu et al., 2026; Ye et al., 2026; Yoon et al., 2026) acts on the mass side: it decides which tokens the per-token loss emphasizes, but still supervises the output rather than the evidence. Attention alignment (Li et al., 2026a; Kim et al., 2026) acts on the placement side: it matches per-token attention maps, each normalized over positions and weighted equally, so a map records where a token looked but not how much it relied on the image, and a function word counts as much as a number. SWIS matches the evidence itself, what each patch contributes to each token, on both sides in one divergence (Eq. (7)). Combining the two lines would need a coefficient between them; here Prop. 1 fixes the balance between mass and placement.

4 Experiments

4.1 Experimental Setup

Models and training data. We train Qwen3.5-2B and Qwen3.5-4B (Qwen Team, 2026a) and run the ablations and analyses at 4B. All runs share one training set of 26K Visual-CoT questions (Shao et al., 2024a), Geometry3K (Lu et al., 2021), and 1.9K Vision-OPD-6K examples (Yuan et al., 2026). Detailed descriptions and splits of these datasets can be found in Appendix D. In the OPSD setting, the teacher π_{T} is the student's own backbone given the box coordinates as a privileged textual hint; in the OPD setting, it is a Qwen3.5-9B trained with GRPO (Shao et al., 2024b) on the same data.

Baselines. In the OPSD setting, we compare with vanilla OPSD, Vision-OPD (Yuan et al., 2026), and ViGOS (Wang et al., 2026a). In the OPD setting, all methods distill from the same teacher: vanilla OPD (Lu & Thinking Machines Lab, 2025), Uni-OPD (Hou et al., 2026), VA-OPD (Liu et al., 2026), and VGS (Yoon et al., 2026).

Benchmarks and evaluation. We evaluate on ten benchmarks in three groups: *visual mathematics* (MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024)), *hallucination* (HallusionBench (Guan et al., 2024), POPE (Li et al., 2023)), and *general visual un-*

Table 2: **Ablation of the evidence loss** on Qwen3.5-4B in the OPD setting. Check marks indicate which terms of Prop. 1 are used and whether they form a single divergence.

Method	Evidence Loss			Accuracy				Energy-PG
	Mass	Place.	Joint	Math	Halluc.	General	Avg.	
Base Model	✗	✗	✗	66.96	80.65	71.93	72.18	24.70
Vanilla OPD ($\mathcal{L}_{\text{out}}$ only)	✗	✗	✗	67.34	80.55	72.15	72.39	27.09
+ mass channel	✓	✗	✗	67.83	81.52	72.85	73.08	28.12
+ placement channel	✗	✓	✗	67.53	81.19	72.48	72.74	30.37
+ both, as two losses (tuned)	✓	✓	✗	68.33	82.28	73.44	73.67	31.46
$\mathcal{L}_{\text{ev}}$ only (no $\mathcal{L}_{\text{out}}$)	✓	✓	✓	66.36	80.34	71.62	71.79	30.84
Ours (joint KL($Q_T \| Q_S$))	✓	✓	✓	**68.76**	**83.00**	**74.13**	**74.29**	**32.50**

Table 3: **Target readout (OPD).** Ours with the target built from other readouts (Appendix C).

Readout	Avg.	Energy-PG
Attention, last layer	72.96	28.43
Attention, all layers	73.21	29.02
Attention × ‖value‖	73.57	30.14
Signed, rectified per layer	73.88	31.26
Ours (summed, then rectified)	**74.29**	**32.50**

Table 4: **Teacher privilege (OPSD).** Ours with other privileged inputs to the self-teacher.

Privilege	Avg.	Energy-PG	Leak (‰)
None (plain self-teacher)	72.24	24.93	–
Answer text	73.38	26.35	41.7
Crop of the region	72.31	25.63	–
Random box	71.37	21.84	0.2
Ours (ground-truth box)	**73.98**	**29.88**	0.2

derstanding (MMVP (Tong et al., 2024), MMMU-Pro (Yue et al., 2025), MMMU (Yue et al., 2024), MMStar (Chen et al., 2024), RealWorldQA (xAI, 2024)). All models are evaluated with lmms-eval (Zhang et al., 2025) under greedy decoding; MathVista and MathVerse are scored by GPT-OSS-120B (OpenAI, 2025) as the judge, and the rest by rule-based matching. Detailed benchmark descriptions and splits are provided in Appendix D.1.

4.2 Main Results

SWIS consistently outperforms existing methods under both teachers and at both scales. As shown in Tab. 1, SWIS achieves the best average accuracy in all four settings, with a mean gain of 1.8 points over its vanilla recipe. At 4B, it is the only method that improves on its vanilla recipe on all ten benchmarks, raising the average by 1.3 points with the self-teacher and 1.9 points with the external teacher. The strongest baselines, ViGOS (OPSD) and VA-OPD (OPD), already gate or reweight the next-token loss by a visual signal, yet trail SWIS by 1.2 points on average over the four settings. Both still distill only next-token distributions, not the evidence behind them. With the external teacher at 4B, SWIS recovers 36% of the gap between the base model and the teacher, while vanilla OPD recovers less than 4%. Notably, this teacher never sees the annotated region, so the benefit of SWIS does not depend on giving the teacher the annotation.

4.3 Ablation Study

To assess the contribution of each component of SWIS, we conduct ablation studies on the 4B models: the objective and the readout are ablated in the OPD setting, and the teacher's privilege in the OPSD setting. Variant definitions and per-benchmark results are in Appendix F.

Both terms of the evidence loss are necessary. As shown in Tab. 2, adding either the mass term (I) or the placement term (II) improves over vanilla OPD, and combining the two brings further gains. The two terms play complementary roles: the mass term mainly improves accuracy, while the placement term mainly improves Energy-PG. Moreover, the joint divergence outperforms the two terms used as separate losses with a tuned balance, in which the placement term is averaged uniformly over tokens as in attention alignment. We attribute this to the W_T weighting of Prop. 1, which focuses the placement term on the load-bearing tokens and fixes the balance between the two terms. Finally, removing $\mathcal{L}_{\text{out}}$ preserves most of the Energy-PG gain but drops accuracy below the base model, so next-token supervision remains necessary.

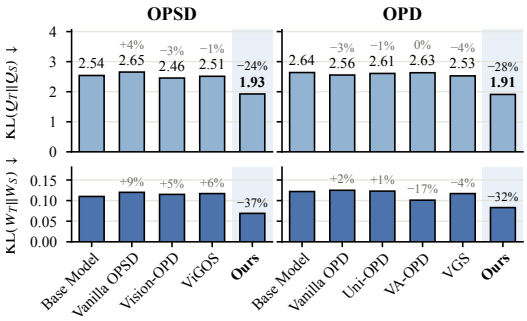


Figure 5: **SWIS narrows the evidence gap; the baselines stay near the base model.** Bottom: its mass term. %: change from the base model.

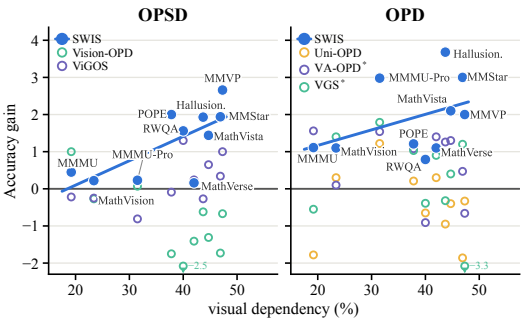
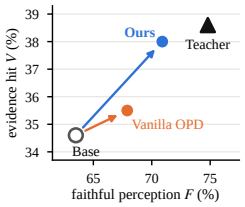


Figure 6: **Accuracy gain over the vanilla baseline grows with visual dependency** (the base 4B model's accuracy drop on a blank image).



Model	Acc.	Process-qualified		$P(A \mid T, V)$			
		PQA↑	UCR↓	T^+V^+	T^+V^-	T^-V^+	T^-V^-
Teacher	64.40	23.27	63.87	85.85	72.40	44.60	31.00
Base Model	57.43	18.00	68.66	82.19	70.33	41.76	29.65
Vanilla OPD	59.70	19.53	67.28	82.89	71.91	40.60	28.40
Ours	**62.90**	**22.53**	**64.18**	84.29	72.91	43.10	29.60

Figure 7: **Vanilla OPD learns what to say; SWIS also learns where to look.** Left: F is the share of responses that describe the image correctly, V the share whose evidence hits the annotated region. Right: PQA counts a correct answer only when its chain is valid (T^+) and its evidence hits the region (V^+); UCR is the share of correct answers that fail either check.

A more faithful readout yields a better target. In Tab. 3, we replace our readout with four alternatives, ordered by how faithfully they attribute a token to the image (Appendix C); Energy-PG is always measured with our readout. Both accuracy and Energy-PG increase along this order. Raw attention, as used in attention distillation (Li et al., 2026a), already improves over vanilla OPD but trails our readout by more than one point of accuracy, suggesting that where the model attends is a weaker target than the evidence it actually uses.

Grounding privilege is the key to transferring where the teacher looks. In Tab. 4, we vary the privileged input of the self-teacher. Without privilege, the teacher is a copy of the initial student, so the student stays close to the base model. The answer text, similar to the reference solution in OPSD (Zhao et al., 2026), improves accuracy but gains far less Energy-PG than the ground-truth box, and its student reproduces the privileged text two orders of magnitude more often (Leak, adapted from the hint-leakage diagnostic of Tian et al. (2026)). A crop of the region, as in Vision-OPD (Yuan et al., 2026), performs close to no privilege. A likely reason is that the answer text and the crop give the teacher what the region shows, so it has less need to read that region in the shared image, whereas the box only tells it where to look. A random box performs worse than no privilege, indicating that the gain comes from the evidence in the target rather than from the extra loss itself.

5 Analysis of SWIS

We next examine whether the gains of SWIS come with the transfer of what the teacher sees.

Analysis 1: SWIS effectively narrows the evidence gap. The evidence loss minimizes the evidence divergence from the teacher. We measure this divergence for every model on held-out responses and, by Prop. 1, split it into the mass term (I) and the placement term (II), the two sides diagnosed in Observations 2 and 3 (Fig. 5). All baselines stay close to the base model, and vanilla OPSD slightly increases the divergence, consistent with Observation 1. VA-OPD, which reweights tokens by visual dependency, reduces only the mass term. SWIS reduces both terms under both teachers, including the external teacher that never sees the annotated region. These results confirm that next-token matching, with or without token reweighting, leaves the evidence gap in place.

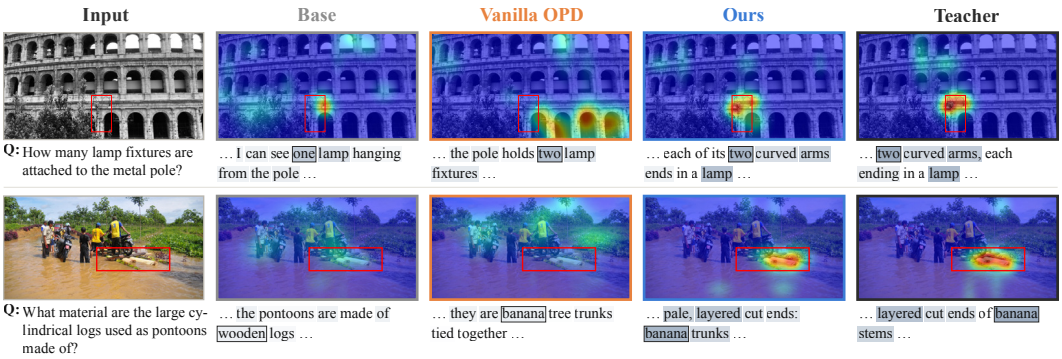


Figure 8: **Vanilla OPD reaches the teacher's answer without the teacher's evidence; SWIS matches both.** Example responses and evidence maps of the base model, vanilla OPD, SWIS, and the teacher. Each map shows the evidence of the boxed word, the fill of each word shows its evidence mass, and the annotated region is outlined in red.

Analysis 2: Vanilla OPD learns what to say, while SWIS also learns where to look. To examine whether this shift in evidence changes how answers are reached, we label each held-out response by four verdicts, following Huang et al. (2025) and Uesato et al. (2022) (Appendix H): the answer is correct (A), the response describes the image correctly (F), its reasoning chain is valid (T), and its evidence hits the annotated region (V). As shown in Fig. 7, vanilla OPD mainly improves what the student says about the image (F) and leaves where it looks (V) nearly unchanged, whereas SWIS improves both and approaches the teacher on V. Notably, accuracy depends far more on the (T, V) cell of a response than on the model within a cell (Fig. 7, right). As a result, SWIS moves both the process-qualified accuracy (PQA) and the unqualified-correct rate (UCR) to near the teacher's level, while vanilla OPD leaves UCR near the base model. Matching the next token transfers what the teacher says; matching the evidence also transfers where it looks.

Analysis 3: The gains of SWIS are larger on benchmarks that depend more on the image. Intuitively, gains that come from better use of the image should grow with how much a benchmark depends on it. Following Chen et al. (2024), we measure this visual dependency as the accuracy that the base 4B model loses when the image is replaced by a blank one (Appendix E). As shown in Fig. 6, the gains of SWIS over its vanilla recipe are positive on every benchmark under both teachers and tend to increase with visual dependency. The least image-dependent benchmarks, MMMU and MathVision, show among the smallest gains, likely because their difficulty lies mainly in knowledge and computation. These results suggest that SWIS helps most where the answer lies in the image.

Analysis 4: Qualitative evaluation of evidence transfer. Fig. 8 shows two examples where vanilla OPD reaches the teacher's answer without the teacher's evidence. Specifically, in the first, it corrects the count to "two" but places the evidence on the arches instead of the lamps (placement, II); in the second, its "banana" draws little evidence, scattered over the people and the vegetation (mass, I). SWIS places the evidence of "two" on both lamps and draws that of "banana" from the logs, and its responses name the same visual cues as the teacher, the curved arms and the layered cut ends. SWIS thus matches the teacher in both respects.

6 Conclusion

In this work, we studied whether on-policy distillation transfers what a multimodal teacher sees along with what it says. We identified an evidence gap: next-token matching aligns the student's outputs with the teacher's, while its visual evidence stays where it started. To address this gap, we proposed SWIS, which matches the student's joint token–patch evidence distribution to the teacher's with a single KL divergence that covers both evidence mass and evidence placement. Experiments on ten benchmarks, two student scales, and two teacher settings show that SWIS consistently outperforms strong OPD and OPSD baselines while narrowing the evidence gap. These results suggest that multimodal distillation benefits from supervising the evidence behind the teacher's predictions, not only the predictions themselves.

AI USE STATEMENT

In this work, we used generative AI tools to annotate the bounding boxes of Geometry3K and to judge model outputs in our evaluation. We have not used generative AI tools to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, to write proofs, to propose or refine hypotheses, to design or give feedback on the research methodology or experiments, to implement our method, to interpret results, or to translate text. Additionally, we used generative AI tools to polish the writing (grammar and wording) and to write and debug plotting and evaluation code. We have reviewed all AI-assisted work: all code was verified and tested by the authors, and all AI-polished text was checked against the authors' intended meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*, 2024.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: Reasoning transfer from LLMs to vision-language models via on-policy distillation. In *CVPR*, 2026.

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? In *NeurIPS*, volume 37, pp. 27056–27087, 2024.

Shizhan Gong, Minda Hu, Qiyuan Zhang, Chen Ma, and Qi Dou. Saliency-R1: Enforcing interpretable and faithful vision-language reasoning via saliency-map alignment reward. In *CVPR*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *ICLR*, 2024.

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. In *CVPR*, pp. 14375–14385, 2024.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs reason in multimodality? EMMA: An enhanced MultiModal ReAsoning benchmark. In *ICML*, volume 267, pp. 22044–22090, 2025.

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

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, et al. Uni-OPD: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, Jingyang Yuan, Wei Ju, Luchen Liu, Tianyu Liu, Baobao Chang, and Ming Zhang. MMEvalPro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. In *NAACL*, pp. 4805–4822, 2025.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *ICLR*, 2026.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. CompoDistill: Attention distillation for compositional reasoning in multimodal LLMs. In *ICLR*, 2026.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *EMNLP*, pp. 7057–7075, 2020.

Bangzheng Li, Jianmo Ni, Chen Qu, Ian Miao, Liu Yang, Xingyu Fu, Muhao Chen, and Derek Zhiyuan Cheng. Reinforced attention learning. *arXiv preprint arXiv:2602.04884*, 2026a.

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 preprint arXiv:2602.02994*, 2026b.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026c.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pp. 292–305, 2023.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Lee Boyd-Graber, Haitao Mi, and Dong Yu. Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization. In *ICLR*, 2026d.

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.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

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

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, pp. 6774–6786, 2021.

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 *ICLR*, 2024.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*, pp. 3428–3448, 2019.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Letitia Parcalabescu and Anette Frank. Do vision & language decoders use images and text equally? How self-consistent are their explanations? In *ICLR*, 2025.

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.

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

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *NeurIPS*, volume 37, pp. 8612–8642, 2024a.

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

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 preprint arXiv:2606.25319*, 2026.

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.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *CVPR*, pp. 9568–9578, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *CVPR Workshops*, 2020.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *NeurIPS*, volume 37, pp. 95095–95169, 2024.

Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation. *arXiv preprint arXiv:2606.19120*, 2026a.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *ICLR*, 2026b.

xAI. Grok-1.5 Vision Preview, April 2024. URL https://x.ai/blog/grok-1.5v. Introduces the RealWorldQA benchmark.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-R1: Advancing multimodal reasoning capabilities of MLLMs via visual perception reward. In *ICLR*, 2026.

Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, Ziming Li, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. Not all tokens see equally: Perception-grounded policy optimization for large vision-language models. *arXiv preprint arXiv:2604.01840*, 2026.

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

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.

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. In *CVPR*, pp. 9556–9567, 2024.

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. In *ACL*, pp. 15134–15186, 2025.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *NAACL Findings*, pp. 881–916, 2025.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, et al. VAD: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MathVerse: Does your multimodal LLM truly see the diagrams in visual math problems? In *ECCV*, 2024.

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*, 2026.

The supplementary material includes a three-minute video overview of the paper (`SWIS_overview_video.mp4`) and its slides. This appendix is organized as follows.

- Appendix A lists the notation.
- Appendix B proves Prop. 1 and places prior objectives on its two terms.
- Appendix C specifies the evidence readout, tests its faithfulness, and checks Observation 1 without it.
- Appendix D gives the data, teachers, prompts, baselines, benchmarks, and hyperparameters.
- Appendix E defines visual dependency and gives the numbers behind Fig. 6.
- Appendix F defines the ablation variants, reports them per benchmark, and varies λ.
- Appendix G reports the evidence divergence of every model and repeats Observations 2 and 3 after training.
- Appendix H gives the judges and prompts of the process evaluation.
- Appendix I discusses the limitations.

A Notation

For ease of reference, Tab. A1 summarizes the notation of the main text and of this appendix. Following the main text, $P(\cdot)$ also denotes a probability over held-out samples, and T also denotes the chain verdict of the process evaluation.

B Proof of Proposition 1

In this section, we first prove Prop. 1, which splits the evidence divergence exactly into a mass term and a placement term (Appendix B.1). We then derive the gradient of $\mathcal{L}_{\text{ev}}$ (Appendix B.2) and show that dependency weighting and attention alignment each act on only one of the two terms (Appendix B.3).

B.1 Proof

We fix one student-generated response $\boldsymbol{y}$ and its image, so that Q_{T} and Q_{S} are distributions over the same $|\boldsymbol{y}| \times J$ token–patch pairs. Let $\tilde{e}(t, j)$ denote the signed sum inside the rectifier of Eq. (2). The teacher's evidence is $e_{\text{T}} = [\tilde{e}_{\text{T}}]_+$, while during training the student's evidence is $e_{\text{S}} = \text{Softplus}_\beta(\tilde{e}_{\text{S}})$ (Sec. 3.3). Accordingly, $W(t) = \sum_j e(t, j)/Z$, $P(j \mid t) = e(t, j)/\sum_{j'} e(t, j')$, and $Q(t, j) = e(t, j)/Z$, where $Z = \sum_{t,j} e(t, j)$ denotes the model's total evidence on the response. We assume $Z_{\text{T}} > 0$; a response on which the teacher places no positive evidence gives no target and is skipped. Three conventions make every term well defined:

(i) $0 \log \frac{0}{b} = 0$ for every $b \geq 0$.

(ii) $P(\cdot \mid t)$ is defined only when $W(t) > 0$. A token with $W_{\text{T}}(t) = 0$ has $Q_{\text{T}}(t, j) = 0$ for every j, so by (i) it contributes nothing to $D_{\text{KL}}(Q_{\text{T}} \| Q_{\text{S}})$ or to $\mathcal{L}_{\text{mass}}$, and the sum in $\mathcal{L}_{\text{place}}$ runs over the tokens with $W_{\text{T}}(t) > 0$.

(iii) Softplus_β is strictly positive, so e_{S}, W_{S}, P_{S}, and Q_{S} are positive on every token and patch.

For either model and every t with $W(t) > 0$,

$$W(t)\, P(j \mid t) = \frac{\sum_{j'} e(t, j')}{Z} \cdot \frac{e(t, j)}{\sum_{j'} e(t, j')} = \frac{e(t, j)}{Z} = Q(t, j), \tag{B.1}$$

Table A1: Notation table for SWIS.

Notation	Description
Task and objective	
$\boldsymbol{I}$, $\boldsymbol{q}$	Input image and question
$\mathcal{D}$	Training set of image–question pairs
$\boldsymbol{y}$, y_t, $\boldsymbol{y}_{<t}$	Response sampled from the student, its t-th token, and the tokens before it
π_{S}, π_{T}	Student and teacher policies
$(\cdot)_{\mathrm{T}}$, $(\cdot)_{\mathrm{S}}$	Computed in the teacher's or the student's forward pass on the same response
$\mathcal{L}_{\mathrm{out}}$	Next-token loss between student and teacher
$\mathcal{L}$, λ	Training objective $\mathcal{L}_{\mathrm{out}} + \lambda\mathcal{L}_{\mathrm{ev}}$, with $\lambda = 0.05$
Evidence readout	
t	Token of the response
j, J	Patch (one visual token) of the image; number of patches
ℓ, h	Softmax-attention layer and head of the backbone
$\alpha_{t,j}^{(\ell,h)}$	Attention weight from the position that predicts y_t to patch j
$\boldsymbol{z}_j^{(\ell-1)}$	Input of layer ℓ at patch j, after the pre-attention normalization
$f^{(\ell,h)}$	Value pathway of head (ℓ, h), $f^{(\ell,h)}(\boldsymbol{z}) = \boldsymbol{z}\,\boldsymbol{W}_{\mathrm{V}}^{(\ell,h)}\boldsymbol{W}_{\mathrm{O}}^{(\ell,h)}$
$\boldsymbol{u}_{y_t}$	Unembedding direction of y_t, with the final normalization folded in
$\tilde{e}(t, j)$	Signed contribution of patch j to the logit of y_t, $\sum_{\ell,h} \alpha_{t,j}^{(\ell,h)} \langle f^{(\ell,h)}(\boldsymbol{z}_j^{(\ell-1)}), \boldsymbol{u}_{y_t}\rangle$
$e(t, j)$	Evidence score, $[\tilde{e}(t, j)]_+$ with $[x]_+ = \max(x, 0)$
$\mathrm{Softplus}_\beta$	Smooth rectifier $\beta^{-1}\log(1 + e^{\beta x})$ of the student branch in training, $\beta = 10$
Evidence distributions and loss	
Z	Total evidence of a response, $\sum_{t,j} e(t, j)$
$W(t)$	Evidence mass: share of the evidence drawn by token t, $\sum_j e(t, j)/Z$
$P(j \mid t)$	Evidence placement: share of token t's evidence that comes from patch j
$Q(t, j)$	Joint evidence distribution, $W(t)\,P(j \mid t) = e(t, j)/Z$
$\mathcal{L}_{\mathrm{mass}}$	Mass term, $D_{\mathrm{KL}}(W_{\mathrm{T}} \,\|\, W_{\mathrm{S}})$
$\mathcal{L}_{\mathrm{place}}$	Placement term, $\sum_t W_{\mathrm{T}}(t)\, D_{\mathrm{KL}}\big(P_{\mathrm{T}}(\cdot \mid t) \,\|\, P_{\mathrm{S}}(\cdot \mid t)\big)$
$\mathcal{L}_{\mathrm{ev}}$	Evidence loss, $D_{\mathrm{KL}}(Q_{\mathrm{T}} \,\|\, Q_{\mathrm{S}}) = \mathcal{L}_{\mathrm{mass}} + \mathcal{L}_{\mathrm{place}}$; the evidence divergence on held-out responses
Evaluation	
b_j	Fraction of patch j that lies inside the annotated box
a	Box-area ratio, $\frac{1}{J}\sum_j b_j$
Energy-PG	Share of a response's evidence inside the box, $\sum_{t,j} Q(t, j)\, b_j$; chance level a
$\mathcal{T}_{\mathrm{LB}}$	Load-bearing tokens: the fewest tokens of a response that hold 60% of W_{T}
In-box share	$\sum_j P(j \mid t)\, b_j$, averaged over $t \in \mathcal{T}_{\mathrm{LB}}$
$d_{\mathrm{out}}(t)$	Next-token divergence from the teacher at token t; converged if below 0.05
VD	Visual dependency: accuracy the base 4B model loses when the image is blanked
A, F, V	Verdicts: answer correct; perception faithful; evidence hits the box, Energy-PG $\geq \tau$
T	Chain verdict: the perception and reasoning judges of Fig. H1 both output Yes
τ	Evidence-hit threshold, $\max(\tau_0,\, c\,a)$ with $c = 2$ and $\tau_0 = 0.15$
PQA	Process-qualified accuracy, $P(A \wedge T \wedge V)$
UCR	Unqualified-correct rate, $1 - \mathrm{PQA}/P(A)$

and both sides vanish when $W(t) = 0$. Substituting $Q_{\mathrm{T}} = W_{\mathrm{T}} P_{\mathrm{T}}$ and $Q_{\mathrm{S}} = W_{\mathrm{S}} P_{\mathrm{S}}$, splitting the logarithm (finite by (iii)), and summing over the tokens with $W_{\mathrm{T}}(t) > 0$ (by (ii)),

$$\begin{aligned} D_{\mathrm{KL}}(Q_{\mathrm{T}} \,\|\, Q_{\mathrm{S}}) &= \sum_t \sum_j W_{\mathrm{T}}(t) P_{\mathrm{T}}(j \mid t) \Big[\log \frac{W_{\mathrm{T}}(t)}{W_{\mathrm{S}}(t)} + \log \frac{P_{\mathrm{T}}(j \mid t)}{P_{\mathrm{S}}(j \mid t)} \Big] \\ &= \sum_t W_{\mathrm{T}}(t) \log \frac{W_{\mathrm{T}}(t)}{W_{\mathrm{S}}(t)} \underbrace{\sum_j P_{\mathrm{T}}(j \mid t)}_{=1} + \sum_t W_{\mathrm{T}}(t) \sum_j P_{\mathrm{T}}(j \mid t) \log \frac{P_{\mathrm{T}}(j \mid t)}{P_{\mathrm{S}}(j \mid t)} \\ &= D_{\mathrm{KL}}(W_{\mathrm{T}} \,\|\, W_{\mathrm{S}}) + \sum_t W_{\mathrm{T}}(t)\, D_{\mathrm{KL}}\big(P_{\mathrm{T}}(\cdot \mid t) \,\|\, P_{\mathrm{S}}(\cdot \mid t)\big) = \mathcal{L}_{\mathrm{mass}} + \mathcal{L}_{\mathrm{place}}, \end{aligned} \tag{B.2}$$

which is exactly Eq. (6). That is, Eq. (B.2) is the chain rule of the KL divergence for a joint distribution written as a marginal times a conditional. Notably, it holds for any positive e_{S}, including the Softplus branch with which the evidence divergences of Fig. 5 are measured. In practice, $\log Q_{\mathrm{S}}$ is computed in log space, with a stable log-Softplus and $\log Z_{\mathrm{S}}$ obtained by log-sum-exp.

B.2 Gradient of the evidence loss

Since Q_{T} is a fixed target, only $Q_{\mathrm{S}}(t, j) = e_{\mathrm{S}}(t, j)/Z_{\mathrm{S}}$ carries gradient. Using $\sum_{t,j} Q_{\mathrm{T}}(t, j) = 1$,

$$\frac{\partial \mathcal{L}_{\mathrm{ev}}}{\partial \log e_{\mathrm{S}}(t, j)} = Q_{\mathrm{S}}(t, j) - Q_{\mathrm{T}}(t, j), \qquad \sum_j \frac{\partial \mathcal{L}_{\mathrm{ev}}}{\partial \log e_{\mathrm{S}}(t, j)} = W_{\mathrm{S}}(t) - W_{\mathrm{T}}(t). \tag{B.3}$$

Specifically, the loss raises the evidence of a token–patch pair where the student has less of it than the teacher, and lowers it where the student has more. Moreover, summing over patches leaves only the mass difference, so a token on which the student draws more evidence than the teacher, such as a function word (Fig. 3b), loses mass through the mass term alone. Finally, through the Softplus branch, the gradient with respect to $\tilde{e}_{\mathrm{S}}(t, j)$ is bounded by $\beta\,|Q_{\mathrm{S}}(t, j) - Q_{\mathrm{T}}(t, j)|$, even where $e_{\mathrm{S}}(t, j) \to 0$.

B.3 Dependency weighting and attention alignment as single-channel cases

As summarized in Tab. B1, the two lines of prior work discussed in Sec. 3.3 each act on only one term of Eq. (6).

Table B1: **Prior objectives act on one term of Eq. (6); SWIS acts on both.**

Family	Methods	Token weight	Placement target	Term of Eq. (6)
Dependency weighting	VPPO (Huang et al., 2026), VA-OPD (Liu et al., 2026), PGPO (Ye et al., 2026), VGS (Yoon et al., 2026)	dependency score	none	mass side only, via weights on $\mathcal{L}_{\mathrm{out}}$
Attention alignment	RAL (Li et al., 2026a), CompoDistill (Kim et al., 2026)	uniform, $1/\lvert\boldsymbol{y}\rvert$	attention map	placement only (Eq. (B.4))
SWIS	Eq. (7)	$W_{\mathrm{T}}(t)$	evidence map $P_{\mathrm{T}}(\cdot \mid t)$	both

Mass side. Setting $P_{\mathrm{S}}(\cdot \mid t) = P_{\mathrm{T}}(\cdot \mid t)$ zeroes the placement term and leaves $D_{\mathrm{KL}}(W_{\mathrm{T}} \| W_{\mathrm{S}})$, which constrains how much evidence each token draws but not where it lands. Dependency weighting acts on this side from outside the divergence. Specifically, it scores each token by how much the image changes its likelihood and reweights $\mathcal{L}_{\mathrm{out}}$ accordingly, leaving the placement free, as Fig. 5 shows for VA-OPD.

Placement side. Let $\bar{\alpha}_{t,j} = \sum_{\ell,h} \alpha_{t,j}^{(\ell,h)}$ and let $P^{\alpha}(j \mid t) = \bar{\alpha}_{t,j} / \sum_{j'} \bar{\alpha}_{t,j'}$ denote the attention map of token t. Normalizing each token separately, $Q'(t, j) = P^{\alpha}(j \mid t)/\lvert\boldsymbol{y}\rvert$, gives both models the

uniform token marginal, so the mass term vanishes and

$$D_{\mathrm{KL}}(Q'_{\mathrm{T}} \,\|\, Q'_{\mathrm{S}}) = \frac{1}{|\boldsymbol{y}|} \sum_t D_{\mathrm{KL}}\big(P^{\alpha}_{\mathrm{T}}(\cdot \mid t) \,\|\, P^{\alpha}_{\mathrm{S}}(\cdot \mid t)\big), \tag{B.4}$$

which recovers per-token attention alignment up to the choice of divergence (Li et al., 2026a). However, this objective weights a function word as much as a number, and attention records where a token looked rather than whether what it read supports the token (Appendix C.1). In our ablations (Tab. 2), the "+ placement channel" and "two losses" rows use Eq. (B.4) with e in place of $\bar{\alpha}$.

C The Evidence Readout and the Diagnosis

Both the diagnosis of Sec. 3.2 and the target of Eq. (7) rest on the readout of Eq. (2). We first specify how the readout is computed, and then show that the patches it ranks highest carry more of a token's image dependence than those ranked by other readouts (Appendix C.1), and that Observation 1 holds without it (Appendix C.2). Throughout, a patch is one visual token, and removing a patch drops that token from the input.

Layers. Qwen3.5 places one softmax-attention layer after every three Gated DeltaNet layers, which gives $N = 6$ such layers in the 2B model and $N = 8$ in the 4B and 9B models. The sum in Eq. (2) runs over all heads of these layers, since Gated DeltaNet mixes tokens through a recurrent state and forms no explicit attention weights over patches. Each head has a sigmoid output gate $\boldsymbol{g}_t^{(\ell,h)} = \sigma(\boldsymbol{z}_t^{(\ell-1)} \boldsymbol{W}_{\mathrm{G}}^{(\ell,h)})$ computed at the predicting position, so the value pathway becomes $f_t^{(\ell,h)}(\boldsymbol{z}) = (\boldsymbol{z}\, \boldsymbol{W}_{\mathrm{V}}^{(\ell,h)} \odot \boldsymbol{g}_t^{(\ell,h)})\, \boldsymbol{W}_{\mathrm{O}}^{(\ell,h)}$ and the sum over patches stays exact. The unembedding direction folds in the final RMSNorm, $\boldsymbol{u}_{y_t} = (\boldsymbol{\gamma} \odot \boldsymbol{E}_{y_t})/\mathrm{rms}(\boldsymbol{h}_t)$, where $\boldsymbol{E}_{y_t}$ is the unembedding row of y_t, $\boldsymbol{\gamma} = \mathbf{1} + \boldsymbol{w}$ is the gain of the zero-centered RMSNorm with weight $\boldsymbol{w}$, and $\boldsymbol{h}_t$ is the final residual at the predicting position.

C.1 Faithfulness of the readouts

Tab. 3 orders five readouts by how faithfully they attribute a token to the image. To establish this order, we run a deletion test on each readout.

Readouts. Let L denote the last softmax-attention layer. The five readouts are: (i) last-layer attention, $\sum_h \alpha_{t,j}^{(L,h)}$; (ii) all-layer attention, $\sum_{\ell,h} \alpha_{t,j}^{(\ell,h)}$; (iii) attention weighted by the value norm (Kobayashi et al., 2020), $\sum_{\ell,h} \alpha_{t,j}^{(\ell,h)} \|f^{(\ell,h)}(\boldsymbol{z}_j^{(\ell-1)})\|$; (iv) the signed contribution rectified per layer, $\sum_\ell \big[\sum_h \alpha_{t,j}^{(\ell,h)} \langle f^{(\ell,h)}(\boldsymbol{z}_j^{(\ell-1)}), \boldsymbol{u}_{y_t}\rangle\big]_+$; and (v) Eq. (2), which rectifies after the sum over layers and heads.

Protocol. For each model, we take the tokens that hold 60% of its evidence mass under Eq. (2) on its greedy held-out responses, so that every readout is tested on the same tokens. We then remove, for each readout, the top $r\%$ of the patches it ranks, re-score the same response, and report the drop in $\log p(y_t \mid \boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y}_{<t})$ as a share of the drop when all patches are removed. Each model is tested with its own readout, and removing random or lowest-ranked patches serves as a control.

Results. As shown in Fig. C1(a, b), the five readouts keep the order of Tab. 3 at every r, for both the teacher and the student, and the top 15% of patches under Eq. (2) carry more of the teacher's image dependence than the whole annotated region. A second check, on the token side, points the same way: a token's evidence mass under Eq. (2) correlates with the log-likelihood it loses when all patches are removed (Spearman $\rho = 0.44$ for the teacher and 0.39 for the student), more strongly than under the other readouts (0.21–0.38).

C.2 Masking the annotated region

Observation 1 measures the evidence gap with the readout. To verify that the gap is not an artifact of the readout, we also measure it with the input alone.

Protocol. We collect each student's newly correct answers (wrong before training and right after) and a random sample of each teacher's correct answers. We then regenerate each response under

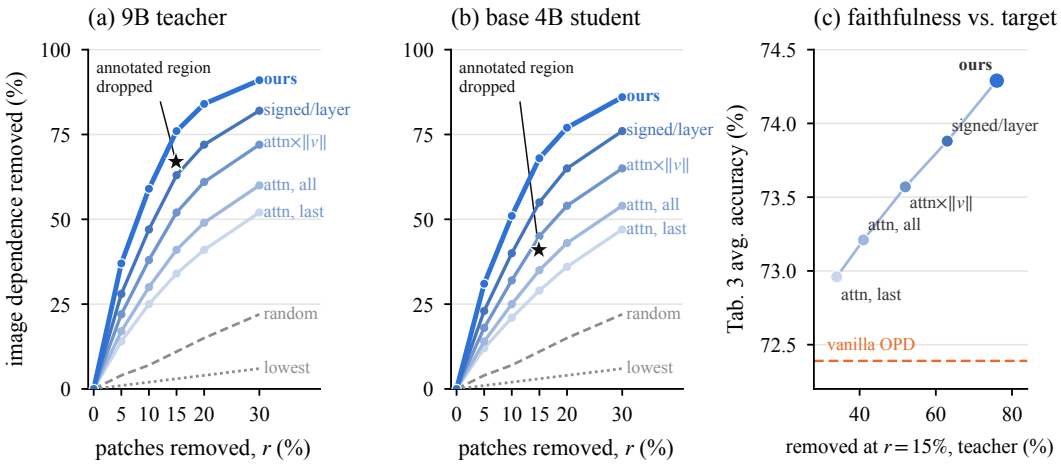


Figure C1: **The readout that removes the most image dependence also builds the best target.** (a, b) Share of a high-mass token's image dependence removed with the top-$r\%$ patches of each readout (star: the annotated region). (c) Share removed at $r = 15\%$ for the teacher against the average accuracy in Tab. 3.

three inputs: (i) with the patches that overlap the annotated region by more than half removed; (ii) with as many patches removed from a random non-overlapping region of the same size (three draws); and (iii) with all patches removed.

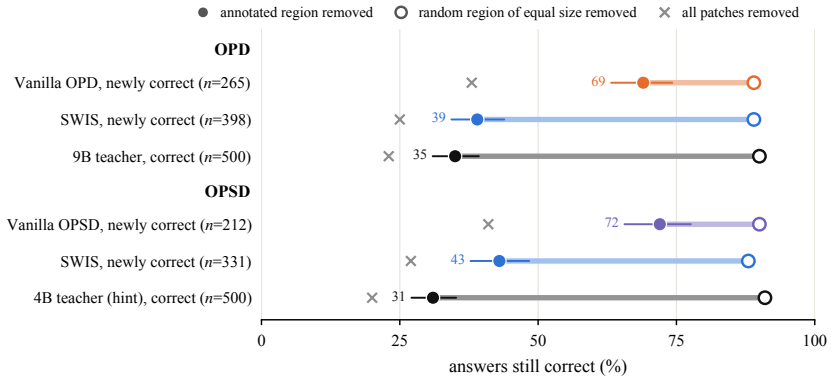


Figure C2: **The answers that next-token matching adds do not rest on the annotated region.** Share still correct after removing the region (filled), a random region of equal size (hollow), or all patches (cross). Qwen3.5-4B.

Results. Most of the vanilla students' new answers survive losing the annotated region, whereas most of the teachers' correct answers do not (Fig. C2); removing a random region of the same size costs every model little. Moreover, part of the new answers survive even without the image. This confirms Observation 1 without the readout. By comparison, the new answers of SWIS depend on the region far more than those of the vanilla students and approach the teachers'. Since this test does not use the readout, it also checks the Energy-PG gains of SWIS independently of the readout that SWIS is trained on.

D Experimental Details

Data. All runs share one training set of about 30K examples from Visual-CoT (Shao et al., 2024a), Geometry3K (Lu et al., 2021), and Vision-OPD-6K (Yuan et al., 2026), each with a short answer and one box around the region that answers the question. For Visual-CoT, Qwen3.5-4B answers each question four times without the box, with the box as a text hint, or with the box drawn on the

image, and GPT-OSS-120B (OpenAI, 2025) judges each answer. The hint raises the mean accuracy from 72.7% to 74.3%, whereas drawing the box lowers it to 59.8%, so the self-teacher receives the box as text. We keep all core examples (0/4 correct without the hint, $\geq$3/4 with it), draw frontier (1/4 without, $\geq$3/4 with) and anchor examples ($\geq$3/4 both ways) evenly across the sources of Visual-CoT, and drop the records with more than one box, which leaves about 26K examples. Geometry3K has no boxes, so Qwen3.8-Max (Qwen Team, 2026b) draws one for each of its 2,101 training problems around the part of the diagram the question refers to. The remaining 1,900 examples are the high-resolution subset of Vision-OPD-6K. We did not remove benchmark images from this training set, which all compared methods share.

Teachers. In the OPSD setting, the teacher is a frozen copy of the initial student that additionally receives the box as text, while the hint never enters the student's input. In the OPD setting, the teacher is Qwen3.5-9B trained with GRPO (Shao et al., 2024b) on the same examples with a binary answer reward and no hint, reaching a held-out Energy-PG of 33.82. The prompts are given in Fig. D1.

Student (both settings) and 9B teacher (OPD)

```
Problem: {q}

Please reason step by step, and put your
final answer within \boxed{}.The output
format should be: <think> reasoning
process here </think> \boxed{}
```

Other privileged inputs (Tab. 4)

None: the student prompt.
Answer text: "Hint: the reference answer is {answer}."
Crop of the region: the cropped region as a second image, with "The second image shows the relevant region."
Random box: the box hint with a box of the same size at a random position that does not overlap the region.

Self-teacher in OPSD, box as text (ours)

The student prompt with the hint after the problem:

```
Hint: the information needed to answer
is in the region [x1, y1, x2, y2] of the
image (pixel coordinates).
```

Red-box input (screening of Visual-CoT)

The box is drawn on the image in red, with "Focus on the region in the red box."

Figure D1: **Prompts.** Privileged inputs enter the teacher's input only; benchmarks use their own prompts.

Implementation. Vision-OPD, ViGOS, and Uni-OPD are retrained on our data from their official code with the recommended hyperparameters. Since VA-OPD and VGS had no public code at the time of writing, we reimplement them following their papers.

D.1 Benchmarks, held-out split, and evidence measures

Benchmarks. All models are evaluated with lmms-eval (Zhang et al., 2025) under greedy decoding and with each benchmark's official prompt (Tab. D1).

Table D1: **Benchmarks.** Judge: GPT-OSS-120B on free-form answers; rule: rule-based matching.

Benchmark	Group	lmms-eval task	Questions	Scoring
MathVista (Lu et al., 2024)	visual math	`mathvista_testmini`	1,000	judge
MathVerse (Zhang et al., 2024)	visual math	`mathverse_testmini`	3,940	judge
MathVision (Wang et al., 2024)	visual math	`mathvision_test`	3,040	rule
HallusionBench (Guan et al., 2024)	hallucination	`hallusion_bench_image`	951	rule
POPE (Li et al., 2023)	hallucination	`pope`	9,000	rule
MMVP (Tong et al., 2024)	general	`mmvp`	300	rule
MMMU-Pro (Yue et al., 2025)	general	`mmmu_pro` (standard)	1,730	rule
MMMU (Yue et al., 2024)	general	`mmmu_val`	900	rule
MMStar (Chen et al., 2024)	general	`mmstar`	1,500	rule
RealWorldQA (xAI, 2024)	general	`realworldqa`	765	rule

Held-out split. For the analyses, we construct a held-out split of 3k examples from Visual-CoT and Vision-OPD-6K that were not selected for training, with a box covering 1–50% of the image, and the source proportions of the training set. Answers are judged by GPT-OSS-120B against the short reference with the prompt in Fig. H1(a), and checked by rule-based matching.

Evidence measures. All evidence measures are defined in Tab. A1. On the held-out split, the box-area ratio a, i.e., the chance level of Energy-PG (Wang et al., 2020), averages 14.9%, and the

load-bearing tokens, ranked by W_{T}, make up 6.8% of all tokens (Fig. 3a). The next-token divergence of token t is defined as $d_{\mathrm{out}}(t) = D_{\mathrm{KL}}(\pi_{\mathrm{S}}(\cdot \mid \boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y}_{<t}) \,\|\, \pi_{\mathrm{T}}(\cdot \mid \boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y}_{<t}))$.

D.2 Hyperparameters and compute

Implementation details. For each prompt, the student samples one response at temperature 1.0, and the teacher stays frozen. $\mathcal{L}_{\mathrm{out}}$ is the reverse KL of Eq. (1), computed over the student's top-100 logits. We set $\lambda = 0.05$, a rollout batch of 512 prompts, and a learning rate of 2×10^{-6}, and train for one epoch on 8 NVIDIA H200 GPUs. Since $\mathcal{L}_{\mathrm{ev}}$ is a single KL over the whole response while $\mathcal{L}_{\mathrm{out}}$ averages over tokens, $\lambda = 0.05$ brings $\lambda\mathcal{L}_{\mathrm{ev}}$ to the scale of $\mathcal{L}_{\mathrm{out}}$ at the start of training; Tab. F2 varies it. Unless otherwise stated, every value is shared across the two scales and the two settings; the other baselines keep their recommended values.

E Visual Dependency

Following the multimodal gain of Chen et al. (2024), we define the visual dependency (VD) of a benchmark as the accuracy that the base 4B model loses when every image is replaced by a blank image of the same size. This keeps the number of patches and the chat template fixed, so that only the image content is removed. The blank-image runs share the evaluation settings of Tab. 1, and Tab. E1 reports the numbers behind Fig. 6.

Table E1: **SWIS gains on every benchmark, more on the more image-dependent half.** Accuracy (%), benchmarks sorted by VD; arrows: gain of SWIS over the vanilla recipe of the same setting.

Benchmark	Base 4B		OPSD setting				OPD setting				
	Acc.	VD	Vanilla	Vision-OPD	ViGOS	SWIS	Vanilla	Uni-OPD	VA-OPD*	VGS*	SWIS
MMMU	68.89	19.2	69.22	**70.22**	69.00	69.67$_{\uparrow 0.45}$	70.11	68.33	**71.67**	69.56	71.22$_{\uparrow 1.11}$
MathVision	47.70	23.4	48.68	48.42	48.42	**48.91**$_{\uparrow 0.23}$	47.50	47.80	47.60	**48.91**	48.59$_{\uparrow 1.09}$
MMMU-Pro	56.82	31.5	57.69	57.76	56.88	**57.92**$_{\uparrow 0.23}$	57.52	58.74	59.06	59.31	**60.50**$_{\uparrow 2.98}$
POPE	87.80	37.8	88.92	87.17	88.83	**90.92**$_{\uparrow 2.00}$	88.49	88.70	89.57	89.52	**89.70**$_{\uparrow 1.21}$
RWQA	76.34	40.0	75.69	73.20	76.99	**77.25**$_{\uparrow 1.56}$	76.86	76.21	75.95	76.47	**77.65**$_{\uparrow 0.79}$
MathVerse	71.19	42.0	72.69	71.29	**72.94**	72.89$_{\uparrow 0.20}$	70.81	71.09	**72.21**	71.70	71.90$_{\uparrow 1.09}$
Hallusion	73.50	43.7	73.98	73.36	73.71	**75.91**$_{\uparrow 1.93}$	72.61	71.66	73.87	72.29	**76.29**$_{\uparrow 3.68}$
MathVista	82.00	44.7	83.50	82.20	84.20	**84.90**$_{\uparrow 1.40}$	83.70	83.30	85.00	84.10	**85.80**$_{\uparrow 2.10}$
MMStar	77.27	46.9	76.13	74.40	76.47	**78.07**$_{\uparrow 1.94}$	76.93	75.07	77.40	78.13	**79.93**$_{\uparrow 3.00}$
MMVP	80.33	47.3	80.67	80.00	81.67	**83.33**$_{\uparrow 2.66}$	79.33	79.00	78.67	76.00	**81.33**$_{\uparrow 2.00}$

As shown in Tab. E1, SWIS gains on all ten benchmarks in both settings, whereas no baseline gains on all ten even in its own setting. Its mean gain on the five more image-dependent benchmarks exceeds that on the other five (1.63 vs. 0.89 points under OPSD, 2.37 vs. 1.44 under OPD). MathVerse is the one exception, with high dependency but a small gain in both settings.

F Ablation Details

We first define the variants of Tabs. 2 to 4, report them per benchmark, and then vary λ (Tab. F2).

Objective (Tab. 2). Starting from vanilla OPD, "+ mass channel" adds $\lambda\mathcal{L}_{\mathrm{mass}}$, and "+ placement channel" adds λ times the placement term with uniform token weights, $\frac{1}{|\boldsymbol{y}|}\sum_t D_{\mathrm{KL}}(P_{\mathrm{T}}(\cdot \mid t) \| P_{\mathrm{S}}(\cdot \mid t))$, as in Eq. (B.4). "+ both, as two losses" adds $\lambda_1\mathcal{L}_{\mathrm{mass}}$ and λ_2 times the uniform placement term, where $\lambda_1 = 0.05$ and the ratio λ_2/λ_1 is tuned over $\{0.25, 0.5, 1, 2, 4\}$. Finally, "$\mathcal{L}_{\mathrm{ev}}$ only" removes $\mathcal{L}_{\mathrm{out}}$.

Readout (Tab. 3). Each row builds Q_{T} and Q_{S} from one of the readouts in Appendix C.1, while Energy-PG is always measured with Eq. (2).

Privilege (Tab. 4). Each row changes only the privileged input of the self-teacher (Fig. D1). Adapting the leak measure of Tian et al. (2026), we define leak as the share of the student's rollouts at the end of training that state the privileged content (box coordinates or reference answer) before their first reasoning step, or refer to a hint that the student never received.

Table F1: **SWIS is best or tied for best on every benchmark in each block.** Per-benchmark results of Tabs. 2 to 4, whose averages and Energy-PG are repeated here; Qwen3.5-4B.

Variant	Visual Math			Hallucination		General					Avg.	Energy-PG
	MathVista	MathVerse	MathVision	Hallusion	POPE	MMVP	MMMU-Pro	MMMU	MMStar	RWQA		
Objective (Tab. 2), OPD setting												
Base Model	82.00	71.19	47.70	73.50	87.80	80.33	56.82	68.89	77.27	76.34	72.18	24.70
Vanilla OPD ($\mathcal{L}_{\text{out}}$ only)	83.70	70.81	47.50	72.61	88.49	79.33	57.52	70.11	76.93	76.86	72.39	27.09
+ mass channel	84.70	71.12	47.66	74.44	88.60	80.33	58.90	70.00	78.40	76.60	73.08	28.12
+ placement channel	83.90	70.99	47.70	73.76	88.62	79.67	58.09	70.11	77.67	76.86	72.74	30.37
+ both, as two losses (tuned)	85.30	71.40	48.29	75.42	89.14	80.67	59.31	70.89	78.93	77.39	73.67	31.46
$\mathcal{L}_{\text{ev}}$ only (no $\mathcal{L}_{\text{out}}$)	81.30	70.96	46.81	72.96	87.72	80.00	57.34	68.56	76.80	75.42	71.79	30.84
Ours (joint KL($Q_{\text{T}}\|Q_{\text{S}}$))	**85.80**	**71.90**	**48.59**	**76.29**	**89.70**	**81.33**	**60.50**	**71.22**	**79.93**	**77.65**	**74.29**	**32.50**
Readout of the target (Tab. 3), OPD setting												
Attention, last layer	84.30	70.69	47.89	73.67	88.75	80.33	58.73	70.67	77.73	76.86	72.96	28.43
Attention, all layers	85.00	71.57	47.96	74.23	88.92	80.00	58.90	70.33	78.07	77.12	73.21	29.02
Attention × ‖value‖	84.80	71.24	48.26	74.58	89.33	80.67	59.83	71.11	78.87	76.99	73.57	30.14
Signed, rectified per layer	85.20	71.65	48.26	75.76	89.42	81.33	59.42	71.00	79.07	77.65	73.88	31.26
Ours (summed, then rectified)	**85.80**	**71.90**	**48.59**	**76.29**	**89.70**	**81.33**	**60.50**	**71.22**	**79.93**	**77.65**	**74.29**	**32.50**
Teacher privilege (Tab. 4), OPSD setting												
Vanilla OPSD (box hint, no $\mathcal{L}_{\text{ev}}$)	83.50	72.69	48.68	73.98	88.92	80.67	57.69	69.22	76.13	75.69	72.72	26.80
None (plain self-teacher)	82.20	71.09	47.70	73.52	87.88	80.67	56.94	69.22	77.13	76.08	72.24	24.93
Answer text	84.50	72.69	48.75	74.75	90.38	82.33	57.80	69.44	76.40	76.73	73.38	26.35
Crop of the region	82.20	71.37	47.96	73.63	87.77	80.67	56.65	69.11	77.13	76.60	72.31	25.63
Random box	81.50	70.43	47.07	71.98	86.45	79.67	56.01	68.44	76.60	75.56	71.37	21.84
Ours (ground-truth box)	**84.90**	**72.89**	**48.91**	**75.91**	**90.92**	**83.33**	**57.92**	**69.67**	**78.07**	**77.25**	**73.98**	**29.88**

As shown in Tab. F1, SWIS is best on every benchmark in all three blocks, tying only with the per-layer rectified readout on MMVP and RWQA, so its advantage does not come from any single benchmark group.

Weight of the evidence loss (Tab. F2). We vary λ in Eq. (8) and keep everything else fixed. Energy-PG rises with λ and approaches the teacher's, while accuracy falls once λ exceeds the default, so a larger weight moves more evidence into the box at some cost in accuracy.

Table F2: **Every λ improves over vanilla OPD, and accuracy peaks at the default.** Weight λ of $\mathcal{L}_{\text{ev}}$ in Eq. (8); Qwen3.5-4B, OPD setting. $\lambda = 0$ is vanilla OPD, and $\lambda = 0.05$ is the default.

λ	0	0.01	0.02	0.05	0.1	0.2
Avg. accuracy	72.39	73.04	73.62	**74.29**	73.91	72.97
Energy-PG	27.09	28.71	30.18	32.50	33.12	**33.38**

G The Evidence Gap after Training

In this section, we examine how training with $\mathcal{L}_{\text{ev}}$ narrows the evidence gap diagnosed in Sec. 3.2. Specifically, we report the evidence divergence of every model (Appendix G.1), repeat Observations 2 and 3 on the trained students (Appendix G.2), and track the gap over training (Appendix G.3).

G.1 Evidence divergence of every model

Tab. G1 reports the numbers behind Fig. 1 (right) and Fig. 5, including the placement term. Each model is measured against the teacher of its setting on its own held-out responses.

The mass term accounts for less than 5% of the divergence for every model (Tab. G1), so the gap lies mostly in placement. SWIS lowers both terms under both teachers. Among the baselines, only VA-OPD lowers the mass term substantially, and none lowers the placement term by even a fifth as much as SWIS.

G.2 Observations 2 and 3 after training

To examine whether SWIS narrows both sides of the gap, we repeat the measurements of Figs. 3a to 3c on the trained students of the OPD setting. As shown in Fig. G1(a, b), under SWIS the evidence mass concentrates on the teacher's load-bearing tokens and moves from function words to numbers and objects. Meanwhile, the in-box share of these tokens rises most of the way to the teacher's (Fig. G1c), including on tokens whose next-token distribution had already converged.

Table G1: **Most of the evidence gap lies in placement, and only SWIS narrows it substantially.** Qwen3.5-4B, held-out split; arrows: change of SWIS from the base model.

Model	Accuracy↑	Energy-PG↑	Evidence divergence from the teacher		
			KL($Q_T \| Q_S$)↓	Mass term↓	Placement term↓
OPSD setting, against the 4B self-teacher					
Base Model	57.43	24.70	2.540	0.110	2.430
Vanilla OPSD	58.73	26.80	2.654	0.120	2.534
Vision-OPD	60.47	26.99	2.455	0.115	2.340
ViGOS	59.97	27.02	2.513	0.117	2.396
SWIS	**61.33**$_{\uparrow 3.90}$	**29.88**$_{\uparrow 5.18}$	**1.925**$_{\downarrow 0.615}$	**0.069**$_{\downarrow 0.041}$	**1.856**$_{\downarrow 0.574}$
4B teacher (hint)	61.63	31.04	–	–	–
OPD setting, against the 9B teacher					
Base Model	57.43	24.70	2.638	0.122	2.516
Vanilla OPD	59.70	27.09	2.555	0.125	2.430
Uni-OPD	59.60	26.88	2.608	0.123	2.485
VA-OPD*	61.17	27.14	2.632	0.101	2.531
VGS*	61.90	26.95	2.527	0.117	2.410
SWIS	**62.90**$_{\uparrow 5.47}$	**32.50**$_{\uparrow 7.80}$	**1.909**$_{\downarrow 0.729}$	**0.083**$_{\downarrow 0.039}$	**1.826**$_{\downarrow 0.690}$
9B teacher	64.40	33.82	–	–	–

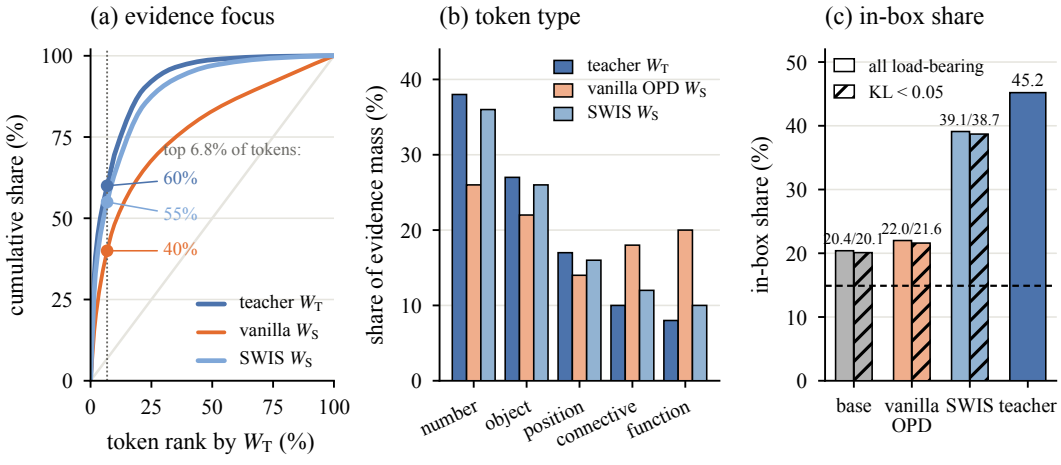


Figure G1: **SWIS narrows both sides of the gap.** (a) Cumulative evidence mass along tokens ranked by W_T; (b) evidence mass by token type; (c) in-box share of the load-bearing tokens, all and with $d_{\text{out}}(t) < 0.05$ (dashed: chance). OPD setting, 4B.

G.3 Training dynamics

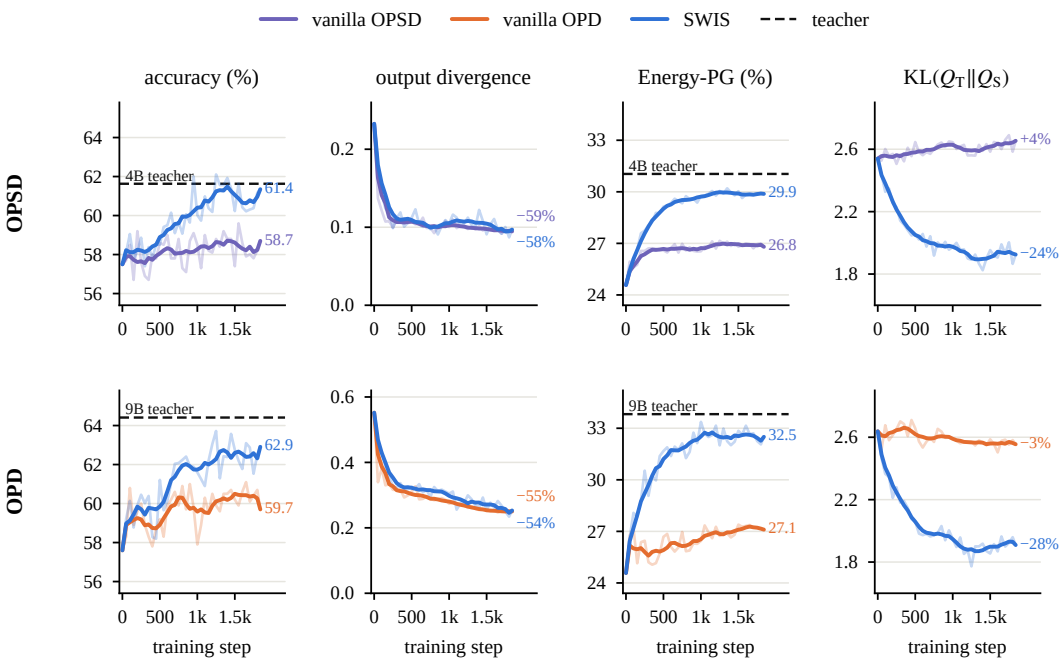


Figure G2: **The evidence moves early under SWIS and barely moves under next-token matching.** The curves of Fig. 2 with SWIS added; OPSD (top) and OPD (bottom), 4B. Faint: one evaluation every 50 steps; solid: moving average over seven evaluations.

The next-token divergence falls at the same rate with and without $\mathcal{L}_{\text{ev}}$ (Fig. G2), so the evidence loss does not slow down next-token matching. Under SWIS, Energy-PG and the evidence divergence move most of the way within the first 500 steps, whereas under the vanilla recipes Energy-PG gains less than three points over the whole run. Moreover, the accuracy of the two students separates after their evidence does.

G.4 Grouping by the teacher's own verdicts

To separate what a student inherits from the teacher, its answers or where it looks, we group the held-out questions by the A and V verdicts of the 9B teacher's own response.

Table G2: **Vanilla OPD inherits the teacher's answers; SWIS inherits where the teacher looks.** Arrows: change from the base model. OPD setting, 4B, held-out split (%).

	Model	Teacher's own response			
		Correct		Wrong	
		In box	Off box	In box	Off box
Accuracy	Base Model	76.34	74.40	29.41	**27.31**
	Vanilla OPD	81.81 $_{\uparrow 5.47}$	77.55 $_{\uparrow 3.15}$	24.55 $_{\downarrow 4.86}$	24.96 $_{\downarrow 2.35}$
	Ours	**85.19** $_{\uparrow 8.85}$	**79.39** $_{\uparrow 4.99}$	**35.81** $_{\uparrow 6.40}$	27.17 $_{\downarrow 0.14}$
Energy-PG	*Teacher (9B)*	50.15	23.49	47.01	20.42
	Base Model	33.80	19.24	32.19	17.06
	Vanilla OPD	36.82 $_{\uparrow 3.02}$	21.21 $_{\uparrow 1.97}$	35.21 $_{\uparrow 3.02}$	18.87 $_{\uparrow 1.81}$
	Ours	**47.71** $_{\uparrow 13.91}$	**22.94** $_{\uparrow 3.70}$	**44.86** $_{\uparrow 12.67}$	**19.99** $_{\uparrow 2.93}$

As shown in Tab. G2, vanilla OPD inherits the answers: its accuracy rises where the teacher is correct and falls below the base model where the teacher is wrong, so next-token matching passes on the teacher's errors as well. SWIS inherits where the teacher looks: its Energy-PG approaches the teacher's where the teacher looks at the box, for right and wrong answers alike, and changes

little elsewhere, even though no model is shown the box. As a result, on the questions the teacher answers wrong while looking at the box, SWIS rises above the base model while vanilla OPD falls below it. These results confirm that the evidence loss transfers where the teacher looks, whether or not the teacher's answer is right.

H Process Evaluation

The process evaluation of Fig. 7 labels each held-out response by its four verdicts. GPT-OSS-120B (OpenAI, 2025) gives the answer verdict A, and Qwen3.8-27B (Qwen Team, 2026b) gives the perception verdict F and the chain verdict T, with the prompts in Fig. H1; T holds only when both the perception and the reasoning judge output Yes. Neither prompt contains the reference answer, and both tell the judge not to judge the final answer, so a chain is not accepted because its conclusion is right. The evidence verdict V holds when the response's Energy-PG reaches $\tau = \max(\tau_0, ca)$, with box-area ratio a, $c = 2$, and $\tau_0 = 0.15$. Following the genuine accuracy of Huang et al. (2025), the process-qualified accuracy is $\mathrm{PQA} = P(A \wedge T \wedge V)$, and the unqualified-correct rate $\mathrm{UCR} = 1 - \mathrm{PQA}/P(A)$ extends the trace-error rate of Uesato et al. (2022) with the evidence verdict. All judges use greedy decoding. The chain judge and every compared model come from the Qwen family, so any preference of the judge for its own family applies to every row alike.

(a) Answer judge for A (GPT-OSS-120B)

```
Your task is to judge whether the response expresses the same meaning as the answer of a
question.
The question is: {question}
The answer is: {gt}
The response is: {response}
Please check and compare them and then judge. If the response is correct, your output
should be Yes. Otherwise, your output should be No. Directly give me your output.
```

(b) Perception judge for F (Qwen3.8-27B; the image is attached)

```
Your task is to judge whether the response describes the image correctly.
The question is: {question}
The key region of the image is: {bbox}
The response is: {response}
Please check every statement in the response about the image (an object, a number, a color,
a position, or a piece of text), using the key region as the main reference. Do not judge
whether the final answer is correct. If every such statement is correct, your output should
be Yes. Otherwise, your output should be No. Directly give me your output.
```

(c) Reasoning judge for T (Qwen3.8-27B; no image)

```
Your task is to judge whether the reasoning in the response is valid.
The question is: {question}
The response is: {response}
Please check every reasoning step in the response (an inference, a calculation, or a
conclusion). Take every statement about the image as given, even if it may be wrong, and
check whether each step follows from the question and the earlier steps. Do not judge
whether the final answer is correct. If every step is valid, your output should be Yes.
Otherwise, your output should be No. Directly give me your output.
```

Figure H1: **Judge prompts.** Only the answer judge sees the reference answer ({gt}); {bbox} is the annotated region as $[x_1, y_1, x_2, y_2]$ in pixels. Each verdict holds when its judge outputs Yes, and T when both (b) and (c) do.

I Limitations

Our work has the following limitations. First, the OPSD teacher uses one annotated region per example, so questions whose evidence is spread over the image gain little from it. Second, Eq. (2) attributes only the direct path from a patch to a token's logit through the softmax-attention layers, so evidence that reaches the token through earlier tokens or through the Gated DeltaNet layers lies outside both the diagnosis and the loss. Third, the target requires the teacher's attention weights and value states, which a teacher behind an API cannot provide. Fourth, all students and teachers come from a single model family, Qwen3.5.