PRESERVE CONTENT, DISRUPT STRUCTURE: PATCH-SHUFFLE COUNTERFACTUAL DISTILLATION FOR VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal on-policy distillation (OPD) provides dense, token-level teacher guidance on student-generated reasoning responses. To identify corrections that depend on visual evidence, previous methods contrast teacher predictions on original inputs and negative views created by masking or downsampling. However, these negatives mainly probe evidence availability and can miss signals for spatial relations and object–attribute bindings. We introduce Patch-Shuffle Counterfactual Distillation (PSCD), which constructs a negative visual view by shuffling patches within the teacher's evidence crop. This preserves pixels within each patch while disrupting spatial and compositional organization. At each student-generated prefix, PSCD compares the frozen teacher's predictions on the intact and shuffled crops, retaining only changes that support the intact teacher's correction of the student to construct a student-anchored distillation target. Experiments on two Qwen3.5 model scales show gains in spatial and compositional reasoning alongside improved fine-grained visual understanding. Among the evaluated Qwen3.5-4B-based methods, PSCD leads on all six spatial benchmarks and outperforms the strongest baseline on all three compositional understanding benchmarks.

1 INTRODUCTION

Multimodal large language models (MLLMs) increasingly answer visual questions through extended reasoning (Bai et al., 2025; Qwen Team, 2026). Effective training must help them interpret and use image evidence within that reasoning process. A reward on the final answer gives limited guidance about which visual decisions need correction (Shao et al., 2024; Yoon et al., 2026). Visual on-policy distillation (OPD) extends the general OPD framework to multimodal reasoning, using teacher predictions to supervise student-generated responses (Yoon et al., 2026). These predictions reflect both image evidence and linguistic context. To improve the student's visual capabilities, we therefore focus on extracting guidance that depends on visual evidence.

Some visual OPD methods give the teacher privileged evidence, such as a crop that makes small image details easier to inspect, while the student answers from the full image (Yuan et al., 2026). To identify guidance that depends on visual evidence, recent methods compare the teacher's predictions on the intact view with those on negative visual views created by masking or downsampling (Liu et al., 2026b; Sun et al., 2026). The question and student prefix remain fixed, so the prediction change probes the teacher's sensitivity to the altered visual input. The choice of negative view therefore determines which aspects of visual dependence the comparison can reveal, making its design important for obtaining useful visual supervision.

We analyze how masking and downsampling change teacher predictions at manually annotated relation and object–attribute binding words in student-generated responses. With the question and response prefix held fixed, downsampling and masking elicit weak teacher responses at 32% and 37% of these positions, respectively, under the criterion detailed in Appendix B.1. Figure 1 illustrates this behavior at the relation word *front* and the attribute word *red*. These observations show that degrading visual evidence can yield little contrastive signal at decisions involving spatial relations and attribute assignments. To obtain additional visual supervision at these decisions, we seek a negative view that preserves local visual content while changing how that content is organized.

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

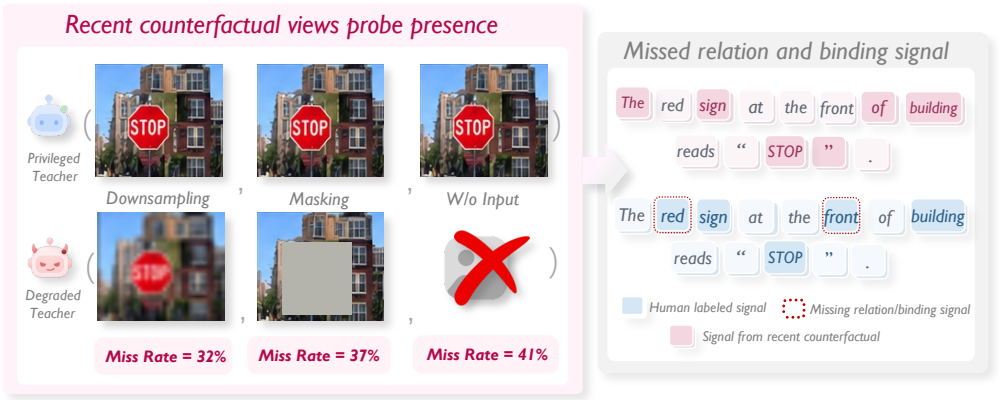

Figure 1: Teacher responses to negative visual views at relation and binding words in student-generated answers. The question and student prefix are fixed while the teacher's visual input changes. The example illustrates weak responses at *front* and *red*.

To this end, we introduce Patch-Shuffle Counterfactual Distillation (PSCD), which constructs this negative view by permuting patches within the teacher's evidence crop. The permutation preserves pixels inside each patch while changing positions and connections across patches. Comparing the frozen teacher's predictions on the intact and shuffled crops at the same student prefix probes its dependence on the original visual arrangement. To turn this response into useful supervision, its direction must support a correction the student needs. PSCD therefore retains candidate updates that agree with the intact teacher's correction and uses them to construct a target around the student's current prediction. The student continues to answer from the full image, learning from the intact crop's visual details and the additional contrast provided by the negative view. By changing spatial and compositional organization, this design also provides a way to supervise decisions about relations and object–attribute bindings.

We evaluate PSCD on Qwen3.5-4B and Qwen3.5-9B against baselines matched in backbone, training data, and post-training budget. The spatial gains cover relative position, perspective-taking, and depth reasoning, while compositional gains reflect better discrimination of attribute assignments and relations between objects (Table 1). For example, the 4B model achieves 51.25% on Winoground, compared with 43.00% for V-Zero, the strongest evaluated baseline at this scale. The benefits also extend to tasks that require locating small regions, recognizing fine details in high-resolution images, and interpreting complex scenes: across six fine-grained benchmarks, the average accuracies reach 77.72% for 4B and 79.20% for 9B, both exceeding the best same-scale baselines (Table 2). Evaluation on MMVP, MMStar, and POPE further yields the highest three-task average at each scale, with improvements over the base models on MMVP and POPE (Table 3).

Our contributions are:

- We analyze teacher responses to common negative visual views in visual OPD and identify weak contrastive signals at annotated relation and object–attribute binding decisions.
- We introduce PSCD, which constructs negative views by shuffling evidence crops and uses agreement with the intact teacher's correction to build a student-anchored distillation target.
- We demonstrate gains in fine-grained visual understanding and spatial and compositional reasoning across two Qwen3.5 model scales, with improved average performance on broader visual tasks.

2 Related Work

On-policy distillation. Knowledge distillation transfers a teacher's predictive distribution to a smaller or otherwise constrained student (Hinton et al., 2015; Kim & Rush, 2016). For autoregressive models, a central difficulty is exposure mismatch between teacher-generated sequences and

the states visited by the student. On-policy approaches instead supervise trajectories sampled from the current student policy (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024; 2025). This principle has also been used to internalize information supplied through a privileged context (Ye et al., 2026). These methods establish student-generated trajectories as an effective basis for dense token-level supervision. Our work adopts the same on-policy setting and studies how the visual view used by the teacher determines the meaning of that supervision.

Privileged visual supervision. Several multimodal OPD methods give the teacher visual information that is easier to use than the student's full image. Vision-OPD transfers the advantage of an evidence-centered crop to a full-image policy (Yuan et al., 2026). VA-OPD identifies tokens whose teacher probabilities depend strongly on fine visual detail and reweights their contribution (Liu et al., 2026b), while ViCuR treats query-relevant visual cues as recoverable privilege (Tian et al., 2026). Other concurrent work gates distillation with contrastive evidence views, separates language and visual grounding gradients, or routes privileged supervision according to token-level advantages (Sun et al., 2026; Yoon et al., 2026; Yu et al., 2026). Together, these studies show that both the source and allocation of privileged supervision matter.

Counterfactual visual signals. Contrasting predictions under altered images has been widely used to expose visual dependence. Visual contrastive decoding and its variants suppress tokens that remain likely under corrupted or less informative visual views (Leng et al., 2024; Chen et al., 2024b; Lee et al., 2024). Differential visual policies similarly compare original, masked, and perturbed inputs to encourage sensitivity to task-relevant evidence while retaining robustness to irrelevant changes (Gao et al., 2026). In multimodal OPD, counterfactual signals have also been used for evidence gating (Sun et al., 2026). These approaches motivate using prediction differences as supervision, but the meaning of that difference still follows from what the intervention changes. Our method makes intervention design the central object of study.

3 Learning from Patch Permutations

Patch-Shuffle Counterfactual Distillation (PSCD) uses an evidence crop to teach the student fine visual details and a patch-permuted copy of that crop to test whether the teacher's answer depends on how those details are arranged. The student generates a response from the full image. At each response position, a frozen teacher evaluates the same prefix under the original and permuted crops. Their difference supplies an arrangement-sensitive signal for distillation. Figure 2 summarizes the training flow.

3.1 A Counterfactual That Changes Patch Order

Let x be the full image, c an evidence crop, and q the question. The student p_θ samples $y \sim p_\theta(\cdot \mid x, q)$. We resize c using the model's vision processor, divide it into nonoverlapping blocks, and randomly permute the blocks over the full crop. For Qwen3.5, Our default uses a 16×16 grid; for a 512×512 processed crop, each shuffled block is therefore 32×32 pixels. The permutation is fixed for each training example. We call the resulting image c^π. Every block in c^π contains the same pixels as in c. Local color, texture, and object parts remain available to the teacher, while their positions and correspondences across blocks change. Consider a red square to the left of a blue circle. Both views retain the colored pieces, while a permutation can break their original pairing and left-right relation. Comparing the teacher's predictions under these views asks whether the original arrangement matters for a candidate answer.

3.2 Turning Patch Order into Supervision

At the student prefix $y_{<t}$, let p_t^S be the student's next-token distribution given (x, q), and let p_t^{ord} and p_t^π be the frozen teacher's distributions given (c, q) and (c^π, q), respectively. All three are evaluated on the same prefix. We restrict them to the student's top-K candidate tokens and a tail bucket, and use $\ell_t(p)$ for centered log probabilities on this shared support. The two changes we need are

$$\Delta_t = \ell_t(p_t^{\mathrm{ord}}) - \ell_t(p_t^\pi), \qquad g_t = \ell_t(p_t^{\mathrm{ord}}) - \ell_t(p_t^S). \tag{1}$$

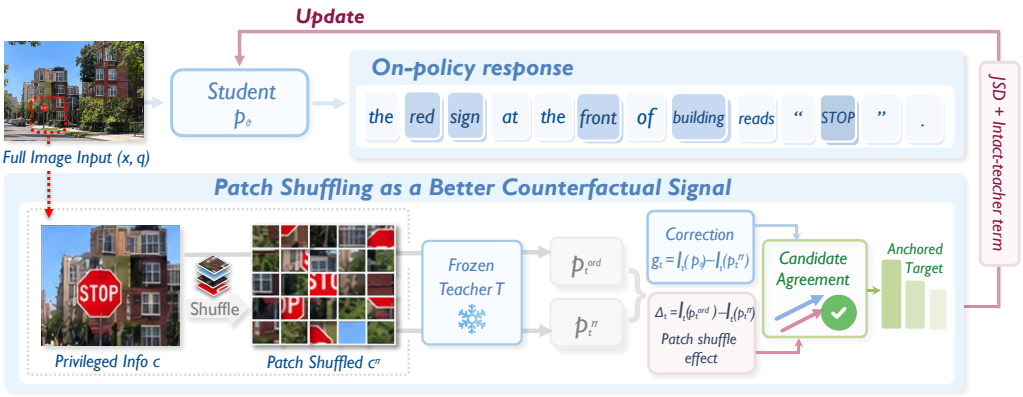

Figure 2: Overview of PSCD. The student responds to the full image, while a frozen teacher evaluates each response prefix on intact and patch-shuffled evidence crops. Patch-order sensitivity is aligned with the teacher's correction to form a student-anchored distillation target.

Here Δ_t measures what restoring the original patch order changes in the teacher's token preferences. The vector g_t measures how the teacher would correct the student with the intact crop. For example, if the teacher favors *left* under the intact crop but not the permuted crop, and it also favors *left* over the student's prediction, that part of its correction is supported by the original arrangement.

We fit a nonnegative multiple of Δ_t to g_t and use the fitted magnitude as an update budget:

$$\lambda_t = \frac{[\langle g_t, \Delta_t \rangle]_+}{\|\Delta_t\|_2^2 + \zeta}, \qquad B_t = \|\lambda_t \Delta_t\|_2. \tag{2}$$

When the two changes disagree overall, $\lambda_t = 0$ and no arrangement-based update is made. Otherwise, we allocate B_t between the candidates that gain and lose support under the intact arrangement. This produces a signed update d_t that can favor *left* while ruling out *right*. Appendix A.1 gives the branch weights.

The budget is allocated at the level of positive and negative candidate groups. Within either group, an individual candidate can still move against the correction proposed by the intact teacher. We keep its update only when the two agree:

$$\tilde{d}_t(v) = d_t(v)\, \mathbf{1}[g_t(v) d_t(v) > 0]. \tag{3}$$

Thus, a patch-order change directly shifts a candidate's target logit only when it supports the intact teacher's correction for that candidate. We form the student-anchored target as

$$q_t^{\mathrm{PS}} = \mathrm{softmax}\Big(\ell_t(p_t^S) + \mathrm{clip}(\tilde{d}_t, -c_0, c_0)\Big). \tag{4}$$

3.3 Learning Objective

We minimize the token-level Jensen–Shannon divergence between the student and q_t^{PS} over valid response positions. An intact-teacher term provides additional guidance when the gated update is small. Its weight a_t decreases as the gated update grows relative to the intact teacher's correction. The objective is

$$\mathcal{L} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \Big[D_{\mathrm{JS}}(\mathrm{stopgrad}(q_t^{\mathrm{PS}}), p_t^S) \quad + \eta\, \mathrm{stopgrad}(a_t) D_{\mathrm{JS}}(\mathrm{stopgrad}(p_t^{\mathrm{ord}}), p_t^S) \Big]. \tag{5}$$

All target-side quantities are detached during optimization. The student continues to learn on its own generated prefixes and answers from the full image; the two crop views are used to construct its training target. Figure 3 schematically compares where Vision-OPD and PSCD place stronger token-level distillation emphasis in three responses.

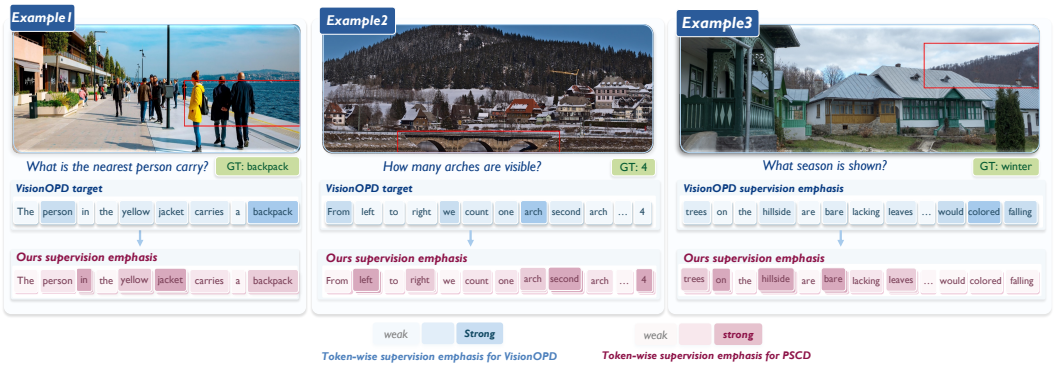

Figure 3: Comparison of token-level distillation emphasis for Vision-OPD and PSCD on three visual questions. Each colored box marks one response position; darker shading qualitatively indicates a stronger training signal at that position. For PSCD, the corresponding loss is the token-level JSD term in Eq. 5. Red boxes identify image regions relevant to the person–backpack binding (left), the arch count (middle), and bare trees as evidence of winter (right).

4 Experiments

4.1 Experimental Setup

Implementation details. We follow the experimental setup from Vision-OPD (Yuan et al., 2026) to train both Qwen3.5-4B and Qwen3.5-9B on 6,241 synthetic visual question-answering examples, ensuring zero overlap with all evaluation benchmarks. For each training prompt, we sample eight on-policy responses from the student.

We use a frozen copy of the initial student model as the teacher. PSCD optimizes token-level Jensen–Shannon divergence between the student and the reconstructed patch-shuffle target over the student's top-100 vocabulary choices, with remaining probabilities grouped into a tail bucket. We add a small intact-teacher regularization term as shown in Equation 5. Training hyperparameters are set as follows: batch size of 96, 8 rollouts per prompt, learning rates of 2×10^{-6} and 5×10^{-7} for the 4B and 9B models, respectively, projection stabilizer $\zeta = 10^{-3}$, logit-shift bound $c_0 = 20$, and regularization weight $\eta = 0.1$. The positive-branch cap τ_+ is 0.8 for 4B and 0.7 for 9B.

Baselines. Table 1 compares spatial and compositional results across scale-matched baselines and external reference models. We evaluate GPT-5.2 and GPT-5.4 (OpenAI, 2025a; 2026), Kimi-K2.6 (Moonshot AI, 2026), GLM-4.6V (Zhipu AI, 2026), and Gemini 3 Flash and Gemini 3.1 Pro (Google DeepMind, 2025; 2026) ourselves. Kimi-K2.6 uses `reasoning_effort=none`, and GLM-4.6V runs with thinking disabled. Published spatial reference scores for the remaining models draw on independent evaluations of OmniSpatial (Lu et al., 2026; Chen et al., 2026; Cho et al., 2026), VSR (Huang et al., 2026), DA-2K (Zhou et al., 2026; Wu et al., 2026), and CV-Bench (Liu et al., 2026a; Zhang et al., 2026). Table 2 extends the comparison to fine-grained visual benchmarks. Its upper blocks include representative "thinking with images" agents (Zheng et al., 2026; Zhang et al., 2025a; Hong et al., 2026; Chng et al., 2025), Gemini 3 Flash and Gemini 3.1 Pro (Google DeepMind, 2025; 2026), and GLM-4.6V (Zhipu AI, 2026). Because these external models vary in architecture, training data, and compute resources, they serve as qualitative reference points. For a strictly matched evaluation, all algorithms in the Qwen3.5 block, including GRPO (Shao et al., 2024), Vision-OPD (Yuan et al., 2026), VA-OPD (Liu et al., 2026b), V-Zero (Sun et al., 2026), Decomposed OPD (Yoon et al., 2026), and PSCD, use the same backbone (Qwen Team, 2026).

Benchmarks. We evaluate four groups of visual benchmarks. For spatial reasoning, we use SAT (Ray et al., 2025), OmniSpatial (Jia et al., 2025), VSR (Liu et al., 2023), DA-2K (Yang et al., 2024), CV-Bench 2D/3D (Tong et al., 2024a) to test relative positions, perspective-taking, and relative depth. For compositional understanding, Winoground (Thrush et al., 2022), SugarCrepe++ (Dumpala et al., 2024), and CRPE (Wang et al., 2024a) test whether models bind objects to their attributes and relations. For fine-grained visual understanding, we use V* (Wu & Xie, 2023), ZoomBench (Wei et al., 2026), HRBench at 4K and 8K (Wang et al., 2024b), and MME-RealWorld

in English and Chinese (Zhang et al., 2025b) to test evidence localization, high-resolution detail recognition, and perception in complex real-world scenes. For broader visual capability, we use MMVP (Tong et al., 2024b), MMStar (Chen et al., 2024a), and POPE (Li et al., 2023) benchmarks to assess visual perception, general multimodal reasoning, and object hallucination. The six fine-grained results in Table 2 use the official Vision-OPD inference and accuracy pipeline with GPT-OSS-120B (OpenAI, 2025b) as judge.

4.2 Main Results

Learning from Patch Shuffling Improves Spatial and Compositional Reasoning. Spatial relations and attribute bindings depend on how visual elements fit together, not just whether each element is visible. Patch shuffling provides a direct training signal for this organization. It keeps the pixels inside each patch unchanged, preserving local color, texture, and shape, while moving the patches away from their original positions and breaking the relations between them. Comparing the teacher's predictions on the intact and shuffled views reveals which corrections depend on the original arrangement.

Table 1: Spatial reasoning and compositional understanding results (%). Avg_6 and Avg_3 are arithmetic means over all six spatial and all three compositional benchmarks, respectively; a missing benchmark leaves its group average unavailable. Within each Qwen3.5 scale, **bold** marks the best result, and the second best is underlined. Omni = OmniSpatial; CV-2D/3D = CV-Bench 2D/3D; Wino = Winoground; Sugar++ = SugarCrepe++.

Model	Param.	*Spatial Reasoning*							*Compositional Understanding*			
		SAT	Omni	VSR	DA-2K	CV-2D	CV-3D	Avg_6	Wino	Sugar++	CRPE	Avg_3
Closed-Source Models												
Gemini 3 Flash	–	72.50	59.17	86.50	82.78	82.71	95.94	79.93	86.88	74.03	78.10	79.67
Gemini 3.1 Pro	–	82.50	64.47	86.87	84.71	82.80	96.67	83.00	88.44	74.03	78.29	80.25
GPT-5.2	–	70.33	51.79	75.99	61.51	75.80	86.83	70.38	53.00	50.58	67.75	57.11
GPT-5.4	–	59.67	58.12	78.50	59.04	78.09	82.67	69.35	54.75	58.97	67.26	60.33
Large Open-Source Models												
GLM-4.6V	106B	64.33	43.97	76.45	62.04	69.68	80.75	66.20	50.75	56.97	67.02	58.25
Kimi-K2.6	1T	76.00	54.27	81.55	56.04	74.34	88.17	71.73	47.25	61.09	65.90	58.08
Qwen3.5-4B Scale												
Qwen3.5-4B	4B	60.33	40.83	77.31	73.79	75.10	90.50	69.64	39.00	66.26	73.40	59.55
Vision-OPD	4B	70.67	49.77	80.50	71.37	81.99	88.67	73.83	41.25	62.60	72.83	58.89
V-Zero	4B	73.00	45.73	79.45	73.74	80.32	89.83	73.68	43.00	64.85	72.53	60.13
PSCD (Ours)	4B	**73.67**	**52.71**	**81.00**	**74.81**	**83.80**	**91.08**	**76.18**	**51.25**	**66.32**	**73.97**	**63.85**
Qwen3.5-9B Scale												
Qwen3.5-9B	9B	73.00	44.94	80.36	77.90	79.14	89.92	74.21	48.75	67.79	75.87	64.14
Vision-OPD	9B	69.67	51.79	82.64	80.71	82.61	90.92	76.39	50.25	68.24	74.85	64.45
PSCD (Ours)	9B	**75.67**	**53.49**	**83.87**	**81.33**	**83.94**	**92.42**	**78.45**	**52.33**	**69.82**	**75.96**	**66.04**

Table 1 matches this expectation. At the 4B scale, PSCD leads on every spatial benchmark and on every compositional benchmarks. The pattern holds across relative position, three-dimensional layout, and assignments between objects and attributes.

The key result is the consistency of these gains. Shuffling does not remove the local visual pieces that identify an object. It changes how those pieces fit together. The student can therefore learn which answers depend on a coherent layout, such as a left or right relation, and which depend on the correct pairing of an object and its attribute. The spatial and compositional results are consistent with our hypothesis that preserving local appearance while changing visual organization provides a useful signal for relation and binding reasoning.

Privileged Evidence Crops Extend the Benefits to Fine-Grained Visual Understanding. The intact evidence crop complements the relational signal from patch shuffling by making fine visual details easier for the teacher to resolve. Fine-grained visual reasoning often depends on evidence that occupies only a small part of the full image. During training, the teacher sees a crop that places this region in a clear view, while the student must answer from the original image. The crop reduces

distraction from the surrounding scene. Learning from the teacher's corrections helps the student find and use the same evidence in the full image.

Table 2: Fine-grained visual performance on six benchmarks. Rows above the Qwen3.5 block show external reference models. All evaluated models use the same official inference pipeline. Values are accuracies (%), and Avg_6 is the unweighted mean. For each Qwen3.5 scale, the best result is **bolded** and the second-best is underlined.

Model	Param.	V*	Zoom	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Avg_6
Thinking-with-Images Agentic Models								
DeepEyes	7B	82.63	46.15	75.50	70.00	63.95	62.41	66.77
Thyme-RL	7B	78.47	46.23	78.87	70.87	64.15	61.41	66.67
DeepEyesV2	7B	77.95	46.11	79.75	72.62	64.25	61.89	67.09
SenseNova-MARS	8B	88.42	48.95	85.00	77.25	67.25	65.72	72.10
Closed-Source Models								
Gemini 3 Flash	–	85.39	61.49	87.25	85.00	73.15	71.66	77.32
Gemini 3.1 Pro	–	88.48	62.01	88.12	84.50	73.32	71.81	78.04
GPT-5.2	–	84.59	54.71	84.64	78.48	71.82	67.43	73.61
GPT-5.4	–	85.83	57.94	86.15	78.10	74.16	70.50	75.45
Large Open-Source Models								
GLM-4.6V	106B	83.20	52.59	83.87	80.13	66.23	66.30	72.05
Kimi-K2.6	1T	83.77	54.67	83.62	79.25	69.88	66.81	73.00
Qwen3.5-4B Scale								
Qwen3.5	4B	82.20	48.28	**85.75**	81.38	63.89	63.58	70.85
GRPO	4B	81.68	56.57	76.75	75.62	71.10	68.58	71.72
VA-OPD	4B	90.05	58.11	81.88	78.75	73.29	67.42	74.92
V-Zero	4B	88.48	56.45	84.88	78.75	73.49	70.45	75.42
Vision-OPD	4B	89.53	59.41	81.75	80.12	**74.51**	70.17	75.92
Decomposed OPD	4B	89.80	58.70	82.60	79.50	73.80	67.82	75.37
PSCD (Ours)	4B	**93.72**	**61.07**	84.75	**82.00**	74.12	**70.68**	**77.72**
Qwen3.5-9B Scale								
Qwen3.5	9B	86.39	52.31	85.12	80.88	71.13	67.31	73.86
GRPO	9B	89.10	57.41	85.12	82.75	73.48	69.35	76.20
VA-OPD	9B	86.91	**62.84**	86.12	82.75	71.17	70.20	76.67
V-Zero	9B	89.53	57.04	85.50	83.88	**75.29**	71.52	77.13
Vision-OPD	9B	92.72	59.26	85.13	83.75	71.76	68.65	76.88
Decomposed OPD	9B	91.40	60.20	85.80	83.10	72.70	69.10	77.05
PSCD (Ours)	9B	**94.76**	61.88	**87.63**	**84.75**	74.33	**71.87**	**79.20**

Table 2 supports this explanation across both model scales. Among Qwen3.5-based methods within the same parameter scale, our 4B model achieves superior performance on fine-grained benchmarks including V*, ZoomBench, and HRBench-8K, while maintaining competitive results on HRBench-4K. This advantage scales consistently to the 9B model, which yields substantial improvements over the original pretrained baseline across V*, ZoomBench, and both HRBench resolutions. Jointly, these benchmarks challenge a model's active vision abilities: localizing small, subtle visual cues, inspecting fine-grained region details, and integrating this localized evidence back into global scene reasoning. These consistent empirical gains confirm that supplying the teacher with privileged, evidence crops effectively guides the student to focus on fine-grained details, complementing the spatial and compositional reasoning gains provided by patch-shuffling counterfactual supervision.

Generalization to Broader Visual Tasks. We next ask whether the gains in spatial, compositional, and fine-grained reasoning come at the expense of broader visual ability. We evaluate PSCD on MMVP (Tong et al., 2024b), MMStar (Chen et al., 2024a), and POPE (Li et al., 2023) to examine visual perception, multimodal reasoning, and object hallucination.

Table 3: Broader visual performance across three tasks (%, higher is better). Avg_3 denotes the unweighted mean, and Δ indicates performance relative to the scale-matched Base. For each model scale, the best result is **bolded** and the second-best is underlined.

Model	MMVP	MMStar	POPE	Avg_3	Δ
		Qwen3.5-4B			
Base	60.00	**80.27**	88.22	76.16	–
GRPO	56.00	69.67	89.19	71.62	−4.54
VA-OPD	54.67	76.73	87.77	73.06	−3.11
V-Zero	60.67	76.87	88.40	75.31	−0.85
Vision-OPD	58.67	78.07	88.80	75.18	−0.98
Decomposed OPD	60.67	79.20	88.53	76.13	−0.03
PSCD (Ours)	**63.33**	78.90	**90.10**	**77.44**	**+1.28**
		Qwen3.5-9B			
Base	68.67	**83.80**	88.68	80.38	–
GRPO	66.00	81.40	89.24	78.88	−1.50
VA-OPD	65.33	78.53	89.36	77.74	−2.64
V-Zero	66.67	79.40	87.86	77.98	−2.41
Vision-OPD	66.67	81.20	89.10	78.99	−1.39
Decomposed OPD	67.33	80.10	88.79	78.74	−1.64
PSCD (Ours)	**71.33**	82.10	**90.33**	**81.25**	**+0.87**

Table 3 shows that PSCD achieves the highest average across these tasks at both model sizes, exceeding the corresponding base models. The gains in targeted visual reasoning are therefore accompanied by stronger overall performance on broader visual tasks.

4.3 Ablation Study and Analysis

Counterfactual View Ablation. To isolate the effect of the counterfactual signal, we train the same PSCD recipe with three teacher views: a downsampled, masked, or patch-shuffled evidence crop. All variants retain the intact teacher crop as the positive view and use the full-image student at inference. Table 4 compares their downstream spatial and compositional benchmark averages. With the PSCD training recipe held fixed, patch shuffle achieves the best results among the three counterfactual views. It scores 76.18 on spatial reasoning and 63.85 on compositional understanding, exceeding the strongest alternative by 1.07 and 2.39 points, respectively. The larger compositional gain suggests that disrupting patch arrangement provides useful supervision for visual relations and bindings. These results support patch shuffling as an effective counterfactual signal for PSCD.

Table 4: Counterfactual-view ablation on Qwen3.5-4B. All variants use the same PSCD training recipe; only the counterfactual teacher view changes. Spatial and compositional averages follow Table 1. Scores are percentages; the best results are **bold** and the second best are underlined.

Counterfactual view	Spatial Avg_6	Compositional Avg_3
Downsampling	75.11	61.46
Masking	75.06	60.78
Patch shuffle	**76.18**	**63.85**

Shuffle Strength and Patch Number. We vary the fraction ρ of patches selected for permutation and the patch grid B on the same evidence crops. At $B = 16 \times 16$, we test $\rho \in \{0.3, 0.6, 1.0\}$ by

permuting patches among $\lceil \rho N \rceil$ of the N grid positions, leaving the rest unchanged. At $\rho = 1.0$, we vary $B \in \{8 \times 8, 16 \times 16, 32 \times 32\}$; each crop is resized to dimensions divisible by 32, so denser grids contain more, smaller patches. All five configurations use the same Qwen3.5-4B initialization, frozen teacher, training examples, target construction, budget, and evaluation protocol. The two sweeps share $(\rho, B) = (1.0, 16 \times 16)$ as their reference. Figure 4 illustrates the controls, and Table 5 reports the results.

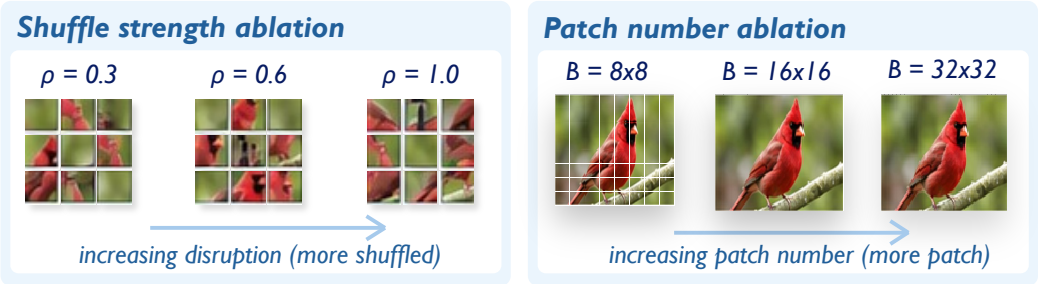

Figure 4: Illustration of shuffle strength and patch number. Left: increasing ρ selects more patches for permutation on a fixed grid. Right: increasing B divides the same crop into more patches; the original crop shows the partitions before permutation.

Table 5: Ablations of shuffle strength and patch number on Qwen3.5-4B. The shared reference is $(\rho, B) = (1.0, 16 \times 16)$. Scores (%) are unweighted averages over the six fine-grained benchmarks in Table 2, six spatial benchmarks and three compositional benchmarks in Table 1. Within each metric, **bold** marks the best result and underlining marks the second best.

Sweep	ρ	B	Fine-grained Avg_6	Spatial Avg_6	Compositional Avg_3
Shuffle strength	0.3	16×16	75.23	74.36	63.17
Shuffle strength	0.6	16×16	76.46	75.78	<u>63.78</u>
Reference	1.0	16×16	<u>77.72</u>	**76.18**	**63.85**
Patch number	1.0	8×8	75.12	75.64	63.24
Patch number	1.0	32×32	**77.83**	<u>76.02</u>	63.62

At a fixed grid, increasing shuffle coverage improves all three task groups, consistent with broader disruption exposing more useful cross-patch dependencies. Grid resolution has a different effect: spatial and compositional performance peak at the intermediate 16×16 partition and then recede at 32×32, while fine-grained performance continues to rise. This pattern suggests that very fine shuffling can fragment texture continuity and scene layout enough to weaken the relation and binding contrasts that the student learns from.

5 Conclusion

We introduced PSCD, a visual OPD method that uses intact and patch-shuffled evidence crops to construct token-level supervision. Patch shuffling preserves patch-local visual content while disrupting spatial relations and object–attribute bindings, allowing the teacher's prediction change to reveal corrections that depend on visual organization. PSCD improves the scale-matched base model in seven of eight comparisons across V⋆, ZoomBench, and HRBench. At the 4B scale, it also achieves the best performance among the evaluated Qwen3.5 methods on all six spatial benchmarks and all three compositional benchmarks. Overall, these results suggest that explicitly perturbing visual organization provides useful supervision beyond simply degrading visual evidence.

6 AI USE STATEMENT

Generative AI tools were used to support literature discovery, discussion of research framing and methodology, organization of experiments, and drafting and editing portions of the manuscript. The authors verified the cited sources, mathematical definitions, experimental results, and all AI-assisted text, and take responsibility for the final content.

7 REPRODUCIBILITY STATEMENT

We will release our code and data once the paper is published. The appendix includes detailed experimental setups and hyperparameters so that others can reproduce our results. We also encourage the community to follow good research practices when using our code and data, to help maintain the reliability and transparency of future work.

8 ETHICS STATEMENT

All datasets used in this work are publicly available and contain no sensitive information.

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, 2024. URL https://arxiv.org/abs/2306.13649.

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

Liang Chen, Haozhe Zhao, Tianyi Ma, and Kuan Li. SWE-Vision: A minimal agent for advancing visual intelligence, 2026. URL https://unipat.ai/blog/SWE-Vision.

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?, 2024a. URL https://arxiv.org/abs/2403.20330.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7824–7846. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/chen24bi.html.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, Gao Huang, Dahua Lin, and Lewei Lu. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning, 2025. URL https://arxiv.org/abs/2512.24330.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, et al. SpatialClaw: Rethinking action interface for agentic spatial reasoning, 2026. URL https://arxiv.org/abs/2606.13673.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. SUGARCREPE++ dataset: Vision-language model sensitivity to semantic and lexical alterations, 2024. URL https://arxiv.org/abs/2406.11171.

Shujian Gao, Yuan Wang, Jiangtao Yan, Zuxuan Wu, and Yu-Gang Jiang. Thinking with deltas: Incentivizing reinforcement learning via differential visual reasoning policy, 2026. URL https://arxiv.org/abs/2601.06801.

Google DeepMind. Gemini 3 Flash: Model card, December 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/. Published December 17, 2025.

Google DeepMind. Gemini 3.1 Pro: Model card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Published February 19, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

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

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model, 2026. URL https://arxiv.org/abs/2511.05271.

YiJie Huang, Yiqun Zhang, Zhuoyue Jia, Xiaocui Yang, Junzhao Huang, Zihan Wang, Shi Feng, Daling Wang, Yifei Zhang, and Yongkang Liu. How many visual tokens do multimodal language models need? scaling visual token pruning with F3A, 2026. URL https://arxiv.org/abs/2605.16359.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models, 2025. URL https://arxiv.org/abs/2506.03135.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. 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 Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 21–27 Jul 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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31044–31062. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Yi-Lun Lee, Yi-Hsuan Tsai, and Wei-Chen Chiu. Delve into visual contrastive decoding for hallucination mitigation of large vision-language models, 2024. URL https://arxiv.org/abs/2412.06775.

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 (CVPR)*, pp. 13872–13882, June 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023. URL https://arxiv.org/abs/2205.00363.

Junze Liu, Kun Qian, Florian Dubost, et al. 3D primitives are a spatial language for VLMs, 2026a. URL https://arxiv.org/abs/2605.12586.

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

Ruijie Lu, Yiyang Ma, Xiaokang Chen, et al. Thinking with visual primitives, 2026. URL https://github.com/xuxinping/Thinking-with-Visual-Primitives/blob/main/Thinking_with_Visual_Primitives.pdf. Technical report.

Moonshot AI. Kimi K2.6: From code to creation, from one to many, 2026. URL https://www.kimi.com/ai-models/kimi-k2-6. Official model page.

OpenAI. Introducing GPT-5.2, December 2025a. URL https://openai.com/index/introducing-gpt-5-2/. Published December 11, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025b. URL https://arxiv.org/abs/2508.10925.

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

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5. Official model release and technical blog.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models, 2025. URL https://arxiv.org/abs/2412.07755.

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

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

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. URL https://arxiv.org/abs/2204.03162.

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

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024a. URL https://arxiv.org/abs/2406.16860.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024b. URL https://arxiv.org/abs/2401.06209.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, Yu Qiao, and Jifeng Dai. The all-seeing project v2: Towards general relation comprehension of the open world, 2024a. URL https://arxiv.org/abs/2402.19474.

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, 2024b. 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, 2026. URL https://arxiv.org/abs/2602.11858.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. SpatioLM: Towards general physical spatial intelligence in vision-language models, 2026. URL https://arxiv.org/abs/2608.01899.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms, 2023. URL https://arxiv.org/abs/2312.14135.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. URL https://arxiv.org/abs/2406.09414.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models, 2026. URL https://arxiv.org/abs/2602.12275.

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, 2026. URL https://arxiv.org/abs/2606.00564.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, Qunzhong Wang, Yang Shi, Xiaobin Hu, Xiangyu Yue, Jiaqi Wang, and Shuicheng Yan. Dopd: Dual on-policy distillation, 2026. URL https://arxiv.org/abs/2606.30626.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation, 2026. URL https://arxiv.org/abs/2605.18740.

Chang-Bin Zhang, Yujie Zhong, Qiang Zhang, and Kai Han. iVGR: Internalizing visually grounded reasoning for MLLMs with reinforcement learning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.31096.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images, 2025a. URL https://arxiv.org/abs/2508.11630.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?, 2025b. URL https://arxiv.org/abs/2408.13257.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning, 2026. URL https://arxiv.org/abs/2505.14362.

Zhipu AI. GLM-4.6V, 2026. URL https://docs.bigmodel.cn/cn/guide/models/vlm/glm-4.6v. Official model documentation.

Yang Zhou, Zixuan Huang, Sunzhu Li, et al. SpatialCLI: Learning to reason with spatial tools, then without them, 2026. URL https://arxiv.org/abs/2607.27703.

APPENDIX

Preserve Content, Disrupt Structure: Patch-Shuffle Counterfactual Distillation for Visual Reasoning

Table of Contents for Appendix

A Implementation Details 15

A.1 Patch-Order Target Construction 15

A.2 Evaluation Protocol Details 15

B Additional Experimental Results 15

B.1 Diagnostic of Negative Visual Views 15

B.2 Inference Efficiency 16

B.3 Effect of the Divergence Objective 16

B.4 Component Ablation 17

B.5 Attention to Object Pairs in Spatial Questions 17

B.6 Qualitative Examples Across Benchmarks 17

A Implementation Details

A.1 Patch-Order Target Construction

The shared support at response position t contains the student's top $K = 100$ tokens and one bucket holding the remaining probability mass. For a distribution p on this support, we use $\ell_t(p) = \log p - \mathrm{mean}_{v \in \mathcal{V}_t} \log p(v)$. Let Δ_t and g_t be the order-restoration effect and intact-teacher correction from Equation 1. The fit and budget are given in Equation 2, where $[z]_+ = \max(z, 0)$ and $\zeta = 10^{-3}$. We split $\Delta_t^+ = [\Delta_t]_+$ and $\Delta_t^- = [\Delta_t]_-$, where $[z]_- = \min(z, 0)$ is the signed negative part. We then measure their agreement with the intact-teacher correction using $s_t^\pm = [\langle g_t, \Delta_t^\pm \rangle]_+$. With $Z_t = s_t^+ + s_t^- + \epsilon$ for $\epsilon > 0$, the branch weights are $w_t^+ = \min(s_t^+ / Z_t, \tau_+)$ and $w_t^- = s_t^- / Z_t$. The signed update is

$$d_t = B_t \left(w_t^+ \frac{\Delta_t^+}{\|\Delta_t^+\|_2 + \epsilon} + w_t^- \frac{\Delta_t^-}{\|\Delta_t^-\|_2 + \epsilon} \right). \tag{6}$$

We then apply the candidatewise agreement rule in Equation 3. The positive cap τ_+ is 0.8 for the 4B model and 0.7 for the 9B model. We clip each coordinate of $\tilde{d}_t$ to $[-c_0, c_0]$ with $c_0 = 20$ before forming the target in Equation 4. Using the pre-clipped update, the weight of the intact-teacher term is $a_t = \mathrm{clip}(1 - \|\tilde{d}_t\|_2 / (\|g_t\|_2 + \epsilon), 0, 1)$, and its coefficient in Equation 5 is $\eta = 0.1$.

A.2 Evaluation Protocol Details

At inference, PSCD receives the original image and the benchmark question. The evidence crop and patch permutation are used only during training; evaluation requires a single model response without test-time crop selection or search. For the Qwen3.5 runs using the Vision-OPD evaluation pipeline, we disable thinking and decode with temperature zero and seed 42.

For the six fine-grained benchmarks in Table 2, we use the Vision-OPD prompts and answer-scoring procedure. The evaluation sets contain 191 V* (Wu & Xie, 2023) questions, 845 ZoomBench (Wei et al., 2026) questions, 800 questions at each HRBench resolution (Wang et al., 2024b), and 23,609 English and 5,917 Chinese MME-RealWorld questions (Zhang et al., 2025b). Answer extraction and rule-based matching are applied first; remaining responses are judged against the reference answer by GPT-OSS-120B (OpenAI, 2025b). We report the percentage of correct questions for each benchmark and the unweighted mean across the six scores, treating the English and Chinese MME-RealWorld sets separately.

Table 1 reports per-benchmark percentages for spatial and compositional reasoning, with CV-Bench 2D and 3D (Tong et al., 2024a) shown separately. Its group averages weight the six spatial and three compositional benchmarks equally. For the broader-task results in Table 3, we score MMVP (Tong et al., 2024b) over 150 question pairs: a pair is correct only if both answers are correct. MMStar (Chen et al., 2024a) uses question-level accuracy. POPE (Li et al., 2023) reports the mean F1 across its adversarial, popular, and random splits. Avg_3 is the unweighted mean of these three benchmark scores.

B Additional Experimental Results

B.1 Diagnostic of Negative Visual Views

The diagnostic in Section 1 measures the frozen teacher's response to visual interventions at annotated relation and object–attribute binding positions in student-generated responses. Let $\mathcal{A}$ denote these positions. At each position t, the teacher scores the student's token y_t under the intact evidence crop c and an altered view $c^{(m)}$, where m denotes the intervention. The question q and student prefix $y_{<t}$ remain fixed. We define the response magnitude and the fraction of annotated positions with weak responses as

$$\begin{aligned} r_t^{(m)} &= \left| \log p_T(y_t \mid q, y_{<t}, c) - \log p_T(y_t \mid q, y_{<t}, c^{(m)}) \right|, \\ \mathrm{MissRate}(m) &= \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \mathbf{1}\left[r_t^{(m)} < \log \tau \right]. \end{aligned} \tag{7}$$

We set $\tau = 1.5$, so a miss means that the intervention changes the teacher's probability of the student's token by less than a factor of 1.5. Under this criterion, downsampling and masking have miss rates of 32% and 37%, respectively, on the annotated positions.

B.2 Inference Efficiency

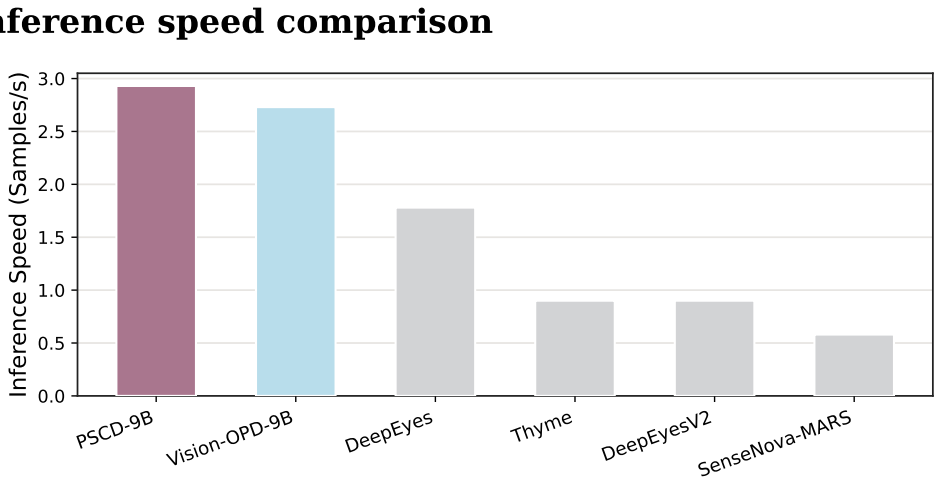

Figure 5: Inference speed comparison on ZoomBench. PSCD-9B achieves the highest inference speed among the compared models while retaining single-pass inference. Baseline rates are reported by Vision-OPD (Yuan et al., 2026).

We compare PSCD-9B with Vision-OPD-9B and representative "Thinking-with-Images" methods, including DeepEyes, Thyme, DeepEyesV2, and SenseNova-MARS, in terms of inference speed. Speed is measured as the inverse of average per-sample inference time on ZoomBench. As shown in Figure 5, PSCD-9B achieves 2.93 samples/s, exceeding Vision-OPD-9B at 2.73 samples/s and the agentic baselines. Unlike agentic and tool-use methods, PSCD answers from the full image in a single model pass. Patch shuffling and the privileged crop teacher are used only during training, so inference requires no iterative crop or search operations.

B.3 Effect of the Divergence Objective

We compare JSD with forward and reverse KL on Qwen3.5-4B, holding the patch-shuffle target, teacher regularization, training data, and compute budget fixed. Forward KL uses the PSCD target as its first argument; reverse KL uses the student distribution. JSD is symmetric. Table 6 evaluates SAT (Ray et al., 2025), OmniSpatial (Jia et al., 2025), and VSR (Liu et al., 2023) for spatial reasoning, and Winoground (Thrush et al., 2022) and SugarCrepe++ (Dumpala et al., 2024) for compositional understanding. Avg_5 is their unweighted mean.

Table 6: JSD versus KL on Qwen3.5-4B.

Benchmark	JSD	F-KL	R-KL
SAT	73.67	73.88	73.92
Omni	52.71	51.12	51.78
VSR	81.00	80.23	80.64
Wino	51.25	50.10	50.62
Sugar++	66.32	65.43	65.78
Avg_5	64.99	64.15	64.55

JSD achieves the highest Avg_5 at 64.99%, exceeding forward KL by 0.84 points and reverse KL by 0.44 points. It leads on OmniSpatial, VSR, Winoground, and SugarCrepe++, while reverse KL is marginally best on SAT (73.92% versus 73.67% for JSD). Reverse KL also exceeds forward KL on all five benchmarks. Under the matched training setup, these results favor JSD for the student-anchored patch-shuffle target.

B.4 Component Ablation

We isolate the projection budget and candidatewise agreement in the patch-shuffle target. The w/o projection variant replaces the projected budget B_t in Equation 2 with $\|\Delta_t\|_2$, retaining the branch allocation and candidatewise agreement. The w/o candidatewise agreement variant uses d_t from Equation 6 directly in the target, without the mask in Equation 3. All other training and evaluation settings match the full Qwen3.5-4B model.

Table 7: Component ablation on Qwen3.5-4B. Scores (%) are unweighted averages over the six fine-grained, six spatial, and three compositional benchmarks in Tables 2 and 1.

Target construction	Fine-grained Avg_6	Spatial Avg_6	Compositional Avg_3
PSCD (full)	77.72	76.18	63.85
w/o projection	77.02	75.32	61.21
w/o candidatewise agreement	77.16	75.24	60.89

Table 7 shows that removing either component reduces performance across all three task groups, with the largest losses in compositional understanding. Projection ties the update magnitude to the intact teacher's correction, so patch-order sensitivity alone cannot set a large training target. Candidatewise agreement further removes token shifts that oppose that correction, even when the overall update is aligned. The larger compositional losses suggest that precise relation and object–attribute binding benefits from both levels of alignment.

B.5 Attention to Object Pairs in Spatial Questions

We compare PSCD with the 4B V-Zero (Sun et al., 2026) and Vision-OPD (Yuan et al., 2026) models on four VStar (Wu & Xie, 2023) questions. Each asks for the left–right relation between two objects. For the answer comparison, we greedily decode from the same full-image prompt. Separately, we probe attention at the final token of a shared "Final answer:" cue.

For each head, we normalize attention over visual tokens before averaging across heads and the last four full-attention layers of Qwen3.5-4B:

$$H(p) = \frac{1}{|\mathcal{L}|N_h} \sum_{\ell \in \mathcal{L}} \sum_{h=1}^{N_h} \frac{A_{q,p}^{(\ell,h)}}{\sum_{p' \in \mathcal{V}} A_{q,p'}^{(\ell,h)}}. \tag{8}$$

Here $\mathcal{L} = \{19, 23, 27, 31\}$, q is the final cue token, and $\mathcal{V}$ contains the visual tokens. We use identical object boxes and a shared heatmap scale across methods within each question.

Figure 6 compares two separated-object questions (A–B) with two questions involving small adjacent objects (C–D). In A, V-Zero and Vision-OPD attend mainly to the telephone booth and answer *left* (B); PSCD also attends to the pillar box and correctly answers *right* (A). In C, PSCD highlights both the faucet and bowl and correctly answers *right* (A), while both baselines answer *left* (B). The stroller in B and the bucket–plunger pair in D show the same increase in attention to both question objects. Across A–D, visual tokens overlapping either outlined object box receive 8.65%, 4.71%, 6.28%, and 25.03% of PSCD's visual attention, compared with 3.17%, 2.54%, 2.19%, and 14.06% for the stronger baseline in each case. PSCD also assigns more attention to each box separately. This broader coverage is consistent with the visual binding encouraged by patch shuffling.

B.6 Qualitative Examples Across Benchmarks

Figure 7 shows six held-out questions from the matched 4B evaluation, with two examples each of fine-grained visual understanding, spatial reasoning, and compositional understanding. We selected

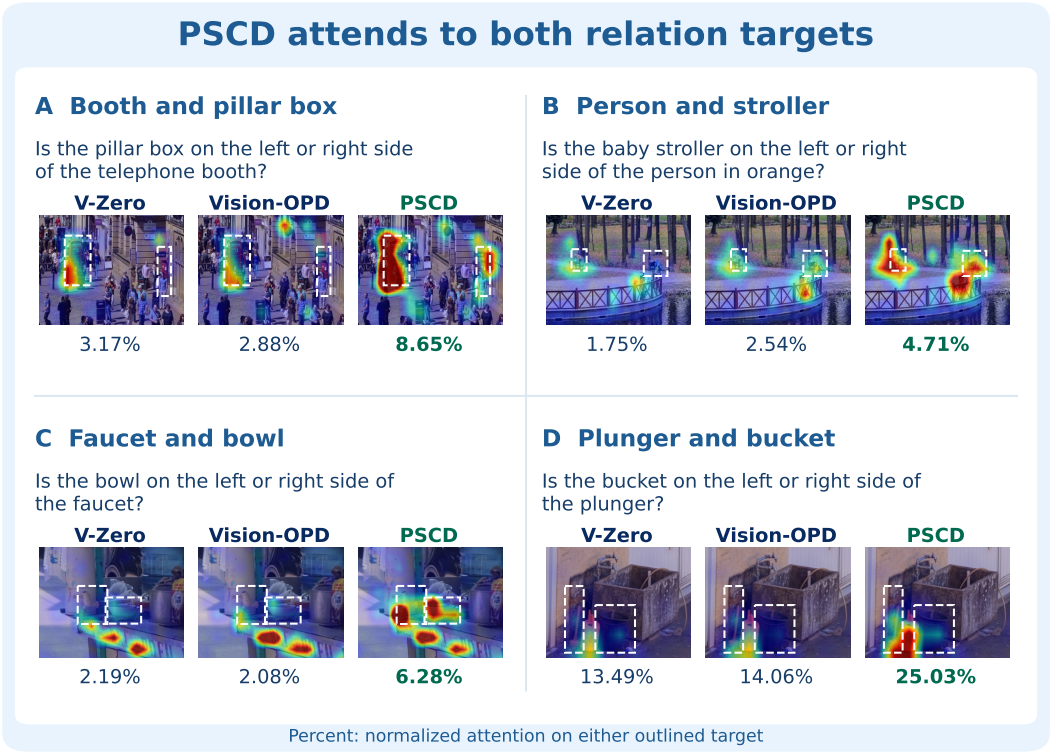

Figure 6: Four VStar spatial-relation questions (samples 154, 120, 152, and 140 for A–D), arranged in two rows. Each panel shows the same image crop under V-Zero, Vision-OPD, and PSCD; every model received the full image. White dashed boxes mark the two question objects. Percentages give normalized attention on visual tokens overlapping either box. Color scales are shared across methods within each panel and clipped at $H(p) = 0.005$ for A–C and $H(p) = 0.04$ for D. PSCD places more combined attention on the target pair in all four cases.

items for which PSCD gives the annotated answer while Qwen3.5-4B-Base, V-Zero-4B, and Vision-OPD-4B give incorrect answers. The displayed answers come from the saved evaluation responses. In V* Bench, PSCD identifies the black straw in a small region of the street image and locates the bowl to the right of the faucet in a cluttered cart. In DA-2K, it selects the closer green point in the foggy cemetery scene; in VSR, it accepts the caption placing the car above the cat. The two compositional examples require distinguishing a wooden spoon *in* a pot from one *above* it, and identifying the child rather than the adult as the reader. These cases illustrate the use of local details and object relations in the final answer.

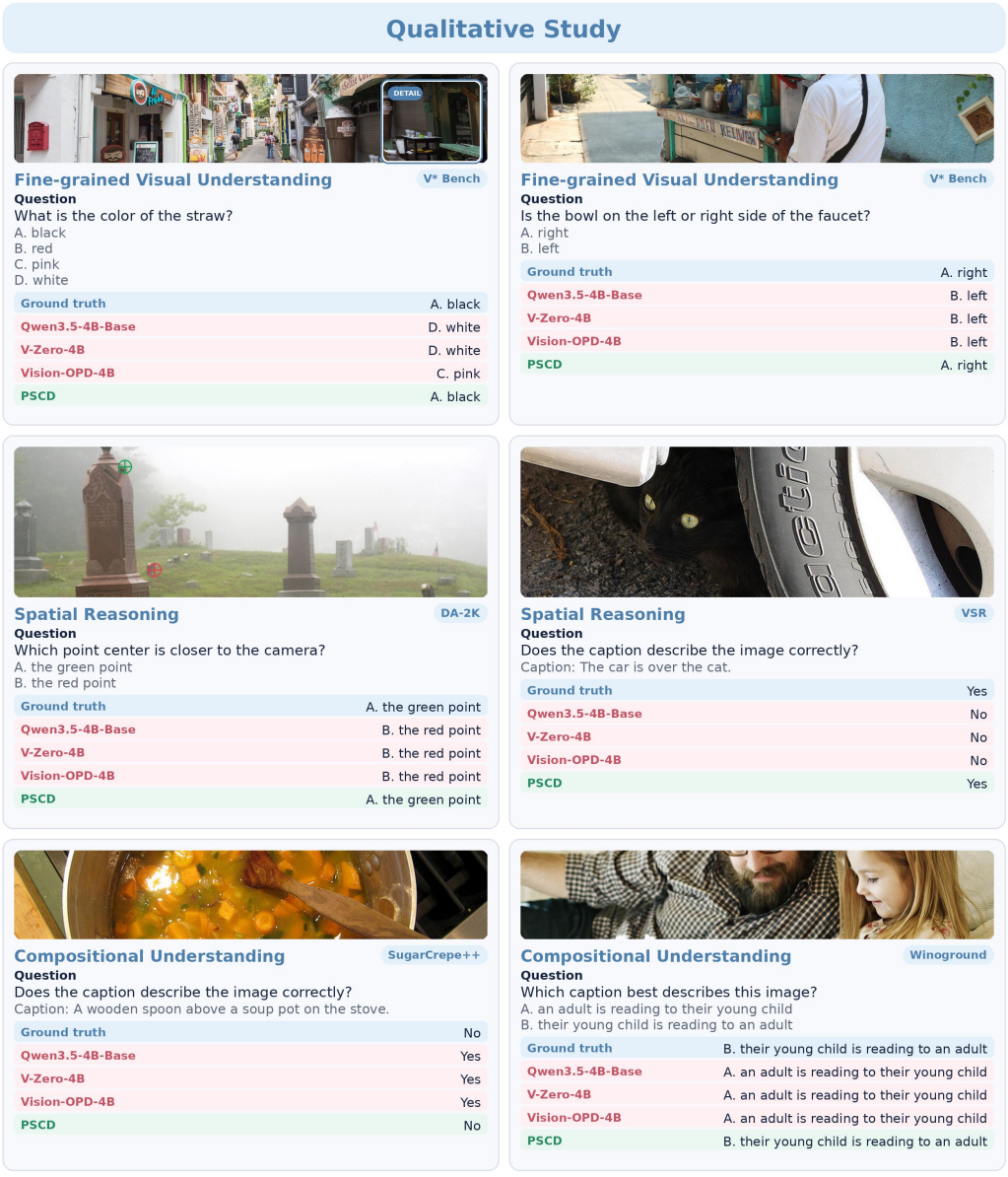

Figure 7: Six examples from the matched 4B evaluation, using PSCD. The three rows show fine-grained visual understanding (V* Bench), spatial reasoning (DA-2K and VSR), and compositional understanding (SugarCrepe++ and Winoground), with two examples in each row. Ground truth and model answers are shown for each item. Red rows are incorrect and green rows are correct. The inset in the first V* Bench panel enlarges the relevant part of the image for display; all models received the full image.