COUNTERFACTUAL VISUAL ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual on-policy distillation uses privileged evidence crops to improve a full-image student's fine-detail recognition. However, distilling teacher outputs on individual images does not explicitly distinguish how the target detail and surrounding content contribute to answer preferences. Our idea is to supplement this supervision with controlled detail changes, observing how teacher preferences respond while holding the question and surrounding scene as constant as possible. We propose CF-OPD, counterfactual visual on-policy distillation, which uses semantic replacement to construct verified image pairs with different correct answers. At each student-generated prefix, CF-OPD compares the two crop-conditioned teacher distributions, using their contrast to adjust a student-anchored distribution before mixing it with the current crop teacher. Independent rollouts on both images let each side learn from this paired supervision. We also introduce CF-Bench-X, an evaluation benchmark comprising 595 verified semantic image pairs. Its Flip rate measures the proportion of pairs answered correctly on both images; its Blind rate measures unchanged predictions despite a verified answer change. CF-OPD achieves the highest average accuracy across six visual benchmarks among the compared Qwen3.5-based methods at both scales. At 9B, it improves recognition and paired evaluation over independent visual OPD on the same inputs. Code and dataset will be released upon acceptance.

1 INTRODUCTION

Recognizing small, answer-determining details is a central challenge for visual question answering on high-resolution images (Wu & Xie, 2024; Wang et al., 2025; Zhang et al., 2024). Visual on-policy distillation (OPD) addresses this challenge by giving a teacher an evidence crop and training a full-image student on its next-token distributions at student-generated prefixes (Yuan et al., 2026; Liu et al., 2026; Tian et al., 2026). This transfers regional perception without additional visual tool calls at inference. However, supervision from an individual image does not explicitly distinguish how the target detail and surrounding content contribute to the teacher's answer preferences. For a red jersey, a teacher distribution favoring "red" supplies the desired answer but does not show how that preference would change if the jersey color changed while the scene remained the same.

Our idea is to use controlled changes to the answer-determining detail to supplement supervision on the current image. Holding the question and surrounding scene as constant as possible makes changes in teacher preferences informative about the replaced attribute. Semantic replacement implements this comparison: replacing the red jersey with a blue one creates two answerable scenes that differ in the detail the student needs to recognize. At a fixed student prefix, the same crop teacher can score both views, expressing its response to the replacement in the token space already used by OPD. The edited image can also be trained on independently through paired SFT or ordinary visual OPD. We investigate whether incorporating the crop comparison into each distillation target improves fine-detail recognition beyond using the same examples independently.

We propose CF-OPD, counterfactual visual on-policy distillation, to implement this paired supervision (Figure 1, right). An automated pipeline plans semantic replacements and routes them to local crop-and-paste or whole-frame editing. Pixel checks, blind answer reading, and a logic audit assess edit quality, the intended answer change, and scene consistency. At each student-generated prefix, the teacher contrast adjusts a student-anchored distribution, which is mixed with the current

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.

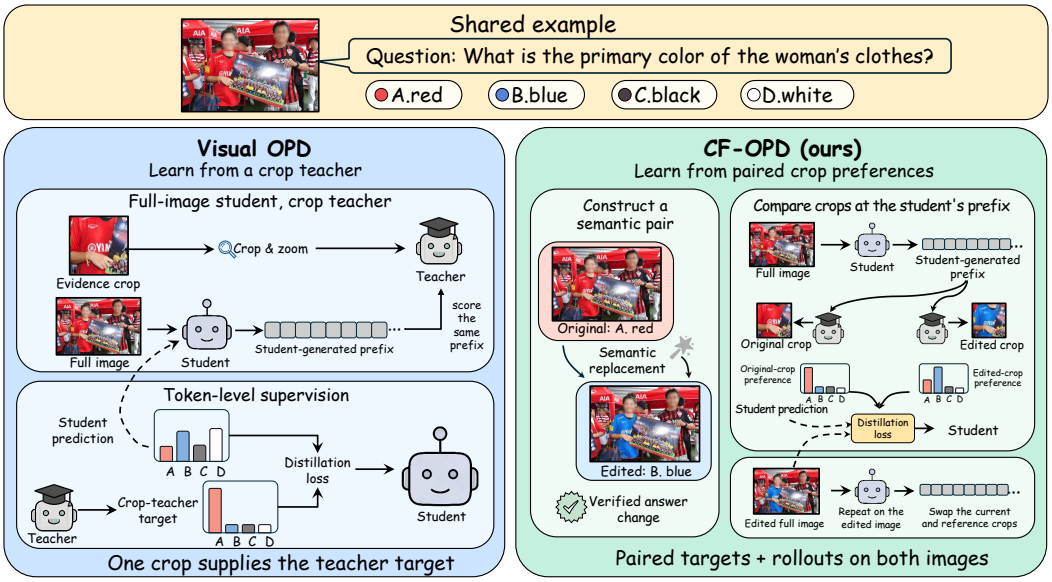


Figure 1: Visual OPD and CF-OPD on a shared example. Left: the student generates a response from the full image, and a crop teacher scores the resulting prefixes to supply the distillation targets. Right: in CF-OPD, a semantic replacement from red to blue with a verified answer change forms a pair. The teacher scores the original and edited crops at the same student prefix, and their comparison forms a paired distillation target. The same procedure runs on the edited image with the two crops swapped.

crop-teacher distribution. The student generates independent rollouts on both images, exchanging the current and reference crops on the edited side. Each side's target therefore uses information from its partner.

We evaluate CF-OPD on six public visual benchmarks to assess fine-grained visual recognition under standard task conditions. CF-Bench-X complements this evaluation with 595 verified image pairs that probe responses to specific edited details. Its flip rate counts pairs answered correctly on both images, and its blind rate counts unchanged answers despite the verified answer change. Together, standard accuracy and paired metrics examine recognition and responses to changed details. Our contributions are:

- **Counterfactual visual on-policy distillation.** We propose CF-OPD, which compares crop-teacher preferences under semantic replacement at each student-generated prefix. It mixes a contrast-adjusted student distribution with the crop teacher and trains on both images, retaining the paired relation within each side's supervision.
- **Verified training-pair construction.** We develop an automated pipeline for semantic edit planning, type-routed image editing, and three-gate verification. Pixel checks, label-blind answer checking, and a logic audit assess edit locality, the intended answer change, and scene consistency to construct training pairs with reliable image-answer relations.
- **Paired evaluation and empirical analysis.** We introduce CF-Bench-X, comprising 595 verified pairs derived from six public visual benchmarks. Standard accuracy, paired-answer metrics, and training comparisons examine fine-detail recognition and the contribution of the paired training signal.

2 Related Work

Visual on-policy distillation. Vision-OPD trains a full-image student with crop-teacher supervision at student-generated prefixes (Yuan et al., 2026). Related methods vary the privileged input, teacher selection, and supervision weighting. ViCuR studies recoverable visual cues (Tian et al., 2026), while S^{2}VOPD augments student views and distills from clean teacher views (Li et al.,

2026). H-OPD selects token-level supervision from heterogeneous teachers (Yin et al., 2026). VA-OPD weights supervision by visual advantage (Liu et al., 2026), and Decomposed OPD separates language-prior and visual-grounding corrections (Yoon et al., 2026). CF-OPD examines how the same crop teacher's candidate preferences respond to a controlled semantic change at a fixed student prefix. In CF-OPD, the crop still provides privileged detail, but the target also depends on a second crop in which that detail has a different value. Both views contribute to each rollout's supervision at the student's current prefix. This preserves the token-level interface of visual OPD while incorporating the teacher's response to the semantic change.

Counterfactual visual supervision. Evidence removal and degradation provide controls for visual supervision. VIGIL penalizes answers that remain confident under visual attention ablation (Xiao et al., 2026), while VCSD erases answer-relevant content and uses the teacher-view difference for self-distillation (Liang et al., 2026). Degraded views also support medical question answering and crop-based visual distillation (Qian et al., 2026; Venkatraman et al., 2026). Evidence-removal contrasts can estimate visually attributed corrections and construct student-anchored targets (Zhang et al., 2026). Semantic replacement supplies two specified values of a detail and a verified answer relation. CF-OPD uses this relation within each side's crop-teacher target, while both images remain valid student inputs. Appendix C discusses additional approaches to teacher privilege and credit assignment.

3 METHOD

3.1 OVERVIEW

CF-OPD uses controlled semantic replacements to provide additional supervision for fine-detail recognition within visual OPD. The student reads a full image and generates its own response; a crop teacher then evaluates the resulting prefixes. A paired crop lets the same teacher evaluate how its candidate-token preferences change when an answer-determining detail takes another value. This comparison uses the same prefixes and token space as the original distillation procedure. Our method has three components: constructing reliable semantic pairs (Section 3.2), converting the crop comparison into a distillation target (Section 3.3), and applying this target to independent rollouts on both images (Section 3.4). Section 3.5 describes the paired evaluation resource.

3.2 CONSTRUCTING VERIFIED SEMANTIC PAIRS

A source example consists of an image x, an evidence box b, a multiple-choice question q, and its answer y. The planner may generate a new question and options; both images in each pair use the same question and options. We seek an edited image x' with a different correct answer $y' \neq y$ to the same question. The edit should change the answer-determining detail while preserving unrelated content as closely as possible. The resulting record (x, x', b, q, y, y') specifies a semantic relation for training. We write $c = \mathrm{crop}(x, b)$ and $c' = \mathrm{crop}(x', b)$ for the teacher views, including the crop padding and resizing used in preprocessing. Figure 2 summarizes the construction procedure.

Semantic edit planning. After sanitizing the evidence box, a planner reads the source example and proposes a plan $\pi = (\tau, d, y', \rho)$. Here τ is the edit type, d describes the intended replacement, and ρ specifies content to preserve. The edit taxonomy includes appearance, written content, object identity, and state or configuration changes. For example, a color plan replaces a red jersey with a blue one while retaining the question and the surrounding scene. Planning constraints discourage additional object changes, reframing, and edits outside the intended region. A self-check assesses whether the proposed edit yields a unique alternative answer and whether the target lies within the evidence region. These checks concern the plan; the rendered image is verified separately.

Type-routed image editing. An editor renders $x' = \mathcal{E}(x, b, \pi)$ using the route selected for the edit type. Local appearance and text edits use a context window around b: the editor modifies the enlarged window, and the result is color-matched and blended back into the source image. Pixels outside that window are preserved before image encoding. Whole-frame editing handles changes such as weather, lighting, or configurations that require broader context, and also provides a fallback

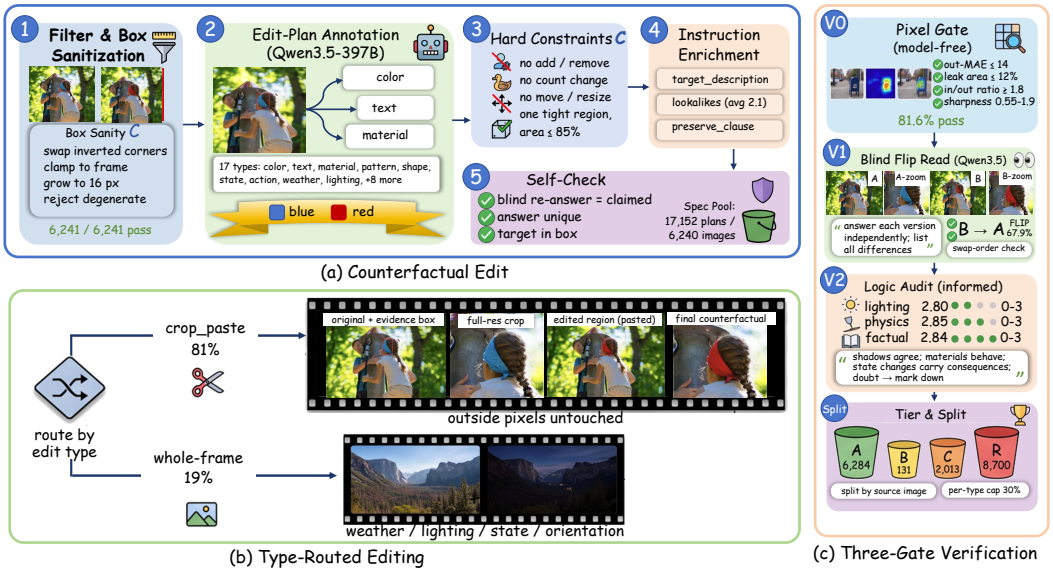


Figure 2: Semantic-pair construction. Edit planning specifies an alternative attribute value and answer. Type-routed editing produces a candidate image. Pixel checks, blind answer reading, and a logic audit assess locality, the answer relation, and scene consistency before quality tiering and source-image splitting.

for large regions. Both routes are subject to the same locality and quality checks. Pixel-based acceptance checks bound unintended changes outside the evidence region.

Three-gate verification. The pixel gate V_0 measures image differences inside and outside b and rejects missing edits, excessive leakage, and large quality shifts. On RGB intensities in $[0, 255]$, the outside mean absolute difference must be at most 14, and at most 12% of outside pixels may have a mean channel difference above 20. The inside mean difference must be at least 6, the inside-to-outside mean difference ratio at least 1.8, and the edited-to-original Laplacian-variance ratio within $[0.55, 1.9]$. The absolute mean brightness shift must be at most 12. These image checks reduce obvious failures before the semantic verification results are used.

The blind answer check V_1 gives a reader the two images, their enlarged evidence crops, and the question, while withholding the edit plan and intended labels. The reader returns a separate answer for each image and lists visible differences. Writing these answers as $\hat{y}$ and $\hat{y}'$, the answer relation passes when

$$\hat{y} = y \quad \text{and} \quad \hat{y}' = y', \qquad y \neq y'. \tag{1}$$

A random subset is also presented in reverse order to audit position bias; inconsistent audited pairs are rejected. The logic audit V_2 receives the edit plan and assesses execution, locality, and lighting, physical, and factual consistency. The three consistency scores range from 0 to 3, and accepted training pairs require each score to be at least 2 together with successful execution and locality checks.

Selection and scope of verification. Pairs that satisfy the answer and quality checks form tiers A and B; tier A additionally requires the highest consistency and execution scores. Quality-controlled candidates without the required answer change enter a weak tier and are excluded from paired training. The accepted pool contains 6,284 tier-A and 131 tier-B pairs. We split by source image and downsample dominant edit types, yielding a pool of 4,659 training pairs and a separate development split. Training-data packing reserves 64 pairs from this pool for validation, leaving 4,595 pairs for optimization. Verification checks the image-answer relation, while the teacher contrast estimates token-level preference changes.

3.3 From Semantic Pairs to Distillation Targets

The visual OPD interface. Let π_θ denote the student and $\bar{\pi}$ the crop teacher, initialized from the same weights and updated by an exponential moving average of the student (Yuan et al., 2026). For a response $\mathbf{y} \sim \pi_\theta(\cdot \mid x, q)$, the student and teacher distributions at position t are

$$p_{S,t} = \pi_\theta(\cdot \mid x, q, \mathbf{y}_{<t}), \qquad p^{+}_{T,t} = \bar{\pi}(\cdot \mid c, q, \mathbf{y}_{<t}). \tag{2}$$

Ordinary visual OPD uses the crop-teacher distribution to supervise these student-generated prefixes. CF-OPD queries the same teacher on c' at the same prefix and evaluates both teacher views on the student's top-K candidate set $\mathcal{V}_t$. Here $p_{S,t}$ and $p^{\pm}_{T,t}$ denote vocabulary probabilities; a tilde denotes their renormalization within $\mathcal{V}_t$.

Teacher contrast at a shared prefix. For each candidate $v \in \mathcal{V}_t$, we compute

$$\begin{aligned} p^{-}_{T,t} &= \bar{\pi}(\cdot \mid c', q, \mathbf{y}_{<t}), \\ \delta_t(v) &= \mathrm{clip}\Big(\log p^{+}_{T,t}(v) - \log p^{-}_{T,t}(v), -\kappa, \kappa\Big). \end{aligned} \tag{3}$$

A positive value means that the teacher assigns the candidate more probability under the original crop than under the edited crop; a negative value indicates the reverse. The question, prefix, teacher parameters, and candidate set are shared between the two evaluations. This makes semantic replacement directly usable within the existing OPD procedure: the comparison produces a token-level signal at a prefix the student actually generated. The contrast is the teacher's estimated response to the crop change. Clipping limits extreme log-ratios from very small teacher probabilities.

Fixed-mixture target. We incorporate this estimate into a student-anchored distribution and mix it with the current crop teacher:

$$q^{*}_t = \mathrm{softmax}_{\mathcal{V}_t}(\log \mathrm{sg}[p_{S,t}] + \alpha\delta_t), \qquad q_t = (1-\lambda)\, q^{*}_t + \lambda\, \tilde{p}^{+}_{T,t}, \tag{4}$$

where sg denotes stop-gradient, α controls the contrast strength, and $\lambda \in [0,1]$ is a fixed teacher-mixture weight. The target is computed at every valid student response prefix. The distribution q^{*}_t uses the semantic contrast, while $\tilde{p}^{+}_{T,t}$ supplies the current crop teacher's preferences. Equation 4 defines both model scales in Table 1 and the 9B reference in Table 2. Appendix B studies a separate gated variant. The contrast adjusts relative candidate preferences according to the log-odds identity in Appendix D.

Distillation loss. The implementation retains a tail bucket for vocabulary mass outside $\mathcal{V}_t$. Let $P_{S,t}$ contain the student's original probabilities on $\mathcal{V}_t$ and tail mass $1 - \sum_{v\in\mathcal{V}_t} p_{S,t}(v)$. Let Q_t contain q_t on $\mathcal{V}_t$ and zero tail mass, with a small numerical floor in implementation. The response loss is

$$\mathcal{L}(x; c, c') = \frac{1}{T}\sum_{t=1}^{T} \omega_t\, \mathrm{JSD}(\mathrm{sg}[Q_t] \,\|\, P_{S,t}), \tag{5}$$

where JSD uses equal mixture weights and ω_t is the detached, clipped importance ratio between the current student and the rollout policy for the sampled token. We mask non-response positions and average the loss over valid response tokens in a minibatch. The per-response notation in Eq. 5 makes the two training sides explicit below. The contrast is computed from vocabulary log-probabilities before top-K renormalization, while target construction normalizes within the candidate set.

3.4 On-Policy Learning on Both Sides

Applying the same rule to the edited image. A verified pair can supervise the student on the original image, on the edited image, or on both. The original side uses x for the student, c for the current-side teacher, and c' as the reference. For the edited side, the student independently samples $\mathbf{y}' \sim \pi_\theta(\cdot \mid x', q)$, and the teacher crop roles are exchanged:

$$\begin{aligned} \delta'_t(v) &= \mathrm{clip}(\log \bar{\pi}(v \mid c', q, \mathbf{y}'_{<t}) - \log \bar{\pi}(v \mid c, q, \mathbf{y}'_{<t}), -\kappa, \kappa), \\ q'^{*}_t &= \mathrm{softmax}_{\mathcal{V}'_t}\big(\log \mathrm{sg}[p'_{S,t}] + \alpha\delta'_t\big). \end{aligned} \tag{6}$$

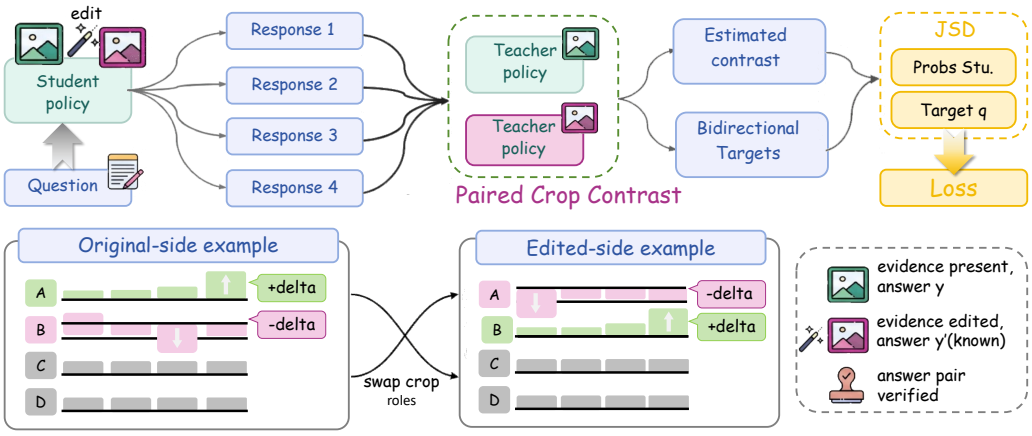


Figure 3: CF-OPD training. The student rolls out independently on each image. At its prefixes, the teacher scores both crops. Their estimated contrast adjusts a student-anchored target, mixed with the current-side teacher. Crop roles swap on the edited side. Bars illustrate possible preference shifts at independent prefixes.

Here $p'_{S,t}$ and $\mathcal{V}'_t$ are the student distribution and candidate set for the edited-side prefix. The same fixed-mixture target and loss use the edited crop as the current-side teacher. Because the rollouts have different prefixes and candidate sets, δ'_t need not equal $-\delta_t$ at corresponding positions. The two examples can be shuffled and optimized independently; each carries both teacher views.

Each training example retains both teacher views, so its supervision uses the pair relation at the student's own prefix. This differs from independently training on the two labeled images and imposes no separate cross-image answer-margin constraint.

Training objective. Let $\mathcal{D}$ contain the training examples with their current and reference crops. For an example without an edit, $c' = c$ and $\delta_t = 0$, so Eq. 4 reduces to a mixture of the detached student and crop-teacher distributions. For a minibatch $\mathcal{B}$ from $\mathcal{D}$ with T_i valid response tokens per example, the token-averaged objective is

$$\mathcal{L}_{\text{train}} = \frac{\sum_{i\in\mathcal{B}} T_i \, \mathcal{L}(x_i; c_i, c'_i)}{\sum_{i\in\mathcal{B}} T_i}, \tag{7}$$

where an edited-side example enters as (x', c', c). Multiple edits can share a source image. At inference, only the full-image student is used; no edited image or crop teacher is required.

3.5 CF-Bench-X: Evaluating Responses to Edited Details

We apply the semantic-editing and verification procedure to questions from six public benchmarks: V*, ZoomBench, HR-Bench 4K and 8K, and MME-RealWorld EN and CN. The resulting CF-Bench-X contains 595 verified pairs and is used only for evaluation; its images, questions, and edits do not enter training. For a model f, both images are answered independently under the same question, and we report

$$\begin{aligned} \text{Flip} &= \frac{1}{N}\sum_i \mathbb{1}[f(x_i, q_i) = y_i \ \wedge \ f(x'_i, q_i) = y'_i], \\ \text{Blind} &= \frac{1}{N}\sum_i \mathbb{1}[f(x_i, q_i) = f(x'_i, q_i)]. \end{aligned} \tag{8}$$

Flip measures correctness on both members of a pair, rather than any answer change. Blind measures unchanged answers despite a verified change in the correct answer; it does not by itself identify the model's internal reason for that behavior. These metrics complement standard visual accuracy by probing the specific detail changes used to motivate paired supervision.

4 EXPERIMENTS

We examine standard visual accuracy, responses to controlled semantic edits, and the contribution of paired supervision beyond independent training on edited examples.

4.1 SETUP

Training. We train Qwen3.5-4B and Qwen3.5-9B (Qwen Team, 2026) using the Vision-OPD source data and verified semantic edits. The teacher starts from the student weights and tracks the student by exponential moving average (Yuan et al., 2026). Training uses minibatches of 96 prompts and 8 rollouts per prompt, with learning rate 2×10^{-6} and top-100 support. Both model scales use the fixed-mixture target in Eq. 4 with paired teacher views. The 9B configuration uses $\alpha = 1$ and $\lambda = 0.15$, starting from the Qwen3.5-9B base checkpoint and training on verified pairs and original training examples. The gated target is analyzed separately in Appendix B. Training runs on one node with eight Cambricon MLU590 accelerators; further ablations and sweeps are in Appendix A.

Benchmarks. Perception is measured on V* (Wu & Xie, 2024), ZoomBench, HR-Bench 4K and 8K (Wang et al., 2025), and MME-RealWorld EN and CN (Zhang et al., 2024), reported as accuracy with Avg_6 the unweighted mean, using greedy decoding and the Vision-OPD inference and scoring scripts, with Qwen3.5-397B as the answer judge for our evaluations; accuracy is always over the full benchmark. Responses to semantic edits are measured on CF-Bench-X (Section 3.5) with Flip and Blind from Eq. 8, where each original and its edit are answered independently and the option letter is parsed.

Baselines. Table 1 lists agentic thinking-with-images models, closed-source models, and large open models for context only, and compares against methods that start from the same Qwen3.5 checkpoints: the base model, GRPO (Shao et al., 2024), and the visual OPD family, Vision-OPD (Yuan et al., 2026), VA-OPD (Liu et al., 2026), V-Zero (Sun et al., 2026), and Decomposed OPD (Yoon et al., 2026). Table 1 includes reported baseline scores for reference; SFT and GRPO on our own data appear in Table 2.

4.2 MAIN RESULTS

Table 1 shows that CF-OPD achieves the best result on most benchmark–scale combinations: seven of the twelve individual comparisons across 4B and 9B. The remaining results are competitive, and CF-OPD obtains the highest Avg_6 among the listed Qwen3.5-based methods at both scales: 78.45 at 9B and 77.41 at 4B. These scores exceed the corresponding base models by 4.59 and 6.56 points, Vision-OPD by 1.57 and 1.49 points, and GRPO by 2.25 and 5.69 points, respectively.

At 9B, CF-OPD leads the same-scale comparison on V*, ZoomBench, and HR-Bench 4K, scoring 93.72, 63.08, and 87.75. Relative to Vision-OPD, the improvements are 1.00 point on V*, 3.82 on ZoomBench, and 2.62 on HR-Bench 4K. On HR-Bench 8K, its 83.50 is comparable to V-Zero's 83.88. On MME-RealWorld, CF-OPD reaches 72.21 in English and 70.46 in Chinese, improving on Vision-OPD by 0.45 and 1.81 points. These competitive MME-RealWorld scores complement its three benchmark wins, with a 1.32-point advantage over V-Zero in Avg_6. Its Avg_6 also exceeds VA-OPD and Decomposed OPD by 1.78 and 1.40 points. Compared with base Qwen3.5-9B, gains occur across all six benchmarks, including 7.33 points on V* and 10.77 on ZoomBench.

At 4B, CF-OPD scores 90.58 on V*, 59.88 on ZoomBench, and 84.62 and 81.50 on HR-Bench 4K and 8K. It leads the same-scale comparison on V*, ZoomBench, HR-Bench 8K, and MME-RealWorld EN. Relative to Vision-OPD, the gains are 1.05 points on V*, 0.47 on ZoomBench, and 2.87 and 1.38 on the two HR-Bench resolutions. The English score of 77.78 is 3.27 points above Vision-OPD, and the Chinese score of 70.10 is comparable to V-Zero's 70.45. CF-OPD also remains competitive on HR-Bench 4K at 84.62, compared with the leading score of 85.75, alongside four benchmark wins and the strongest average. The 4B Avg_6 is also 2.49 points above VA-OPD and 2.04 above Decomposed OPD. Together, the two scales show improved overall recognition, with prominent gains on high-resolution visual benchmarks and no additional inference-time visual tools.

Table 1: Accuracy (%) on six visual benchmarks; Avg_6 is the unweighted mean. Highlighted rows are ours. Within each Qwen3.5 scale, bold and underlined values mark the best and second-best results.

Model	Param.	V*	Zoom	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Avg_6
"Thinking-with-Images" Agentic Models								
DeepEyes (Zheng et al., 2025)	7B	82.63	46.15	75.50	70.00	63.95	62.41	66.77
Thyme-RL (Zhang et al., 2025)	7B	78.47	46.23	78.87	70.87	64.15	61.41	66.67
DeepEyesV2 (Hong et al., 2026)	7B	77.95	46.11	79.75	72.62	64.25	61.89	67.09
SenseNova-MARS (Chng et al., 2025)	8B	88.42	48.95	85.00	77.25	67.25	65.72	72.10
Closed-Source Models								
Gemini 3 Flash (Google DeepMind, 2025)	–	85.39	61.49	87.25	85.00	73.15	71.66	77.32
Gemini 3.1 Pro (Google DeepMind, 2026)	–	88.48	62.01	88.12	84.50	73.32	71.81	78.04
GPT-5.2 (OpenAI, 2025)	–	84.59	54.71	84.64	78.48	71.82	67.43	73.61
GPT-5.4 (OpenAI, 2026)	–	85.83	57.94	86.15	78.10	74.16	70.50	75.45
Large Open-Source Models								
Qwen3-VL-Instruct (Bai et al., 2025)	235B	87.37	57.23	88.00	79.25	71.09	65.86	74.80
Qwen3.5-397B-A17B (Qwen Team, 2026)	397B	84.23	58.30	91.25	84.37	74.17	66.64	76.49
GLM-4.6V (Zhipu AI, 2026)	106B	83.20	52.59	83.87	80.13	66.23	66.30	72.05
Kimi-K2.6 (Moonshot AI, 2026)	1T	83.77	54.67	83.62	79.25	69.88	66.81	73.00
Qwen3.5-Based Models – 4B Scale								
Qwen3.5 (Qwen Team, 2026)	4B	82.20	48.28	**85.75**	81.38	63.89	63.58	70.85
GRPO (Shao et al., 2024)	4B	81.68	56.57	76.75	75.62	71.10	68.58	71.72
VA-OPD (Liu et al., 2026)	4B	90.05	58.11	81.88	78.75	73.29	67.42	74.92
V-Zero (Sun et al., 2026)	4B	88.48	56.45	84.88	78.75	73.49	**70.45**	75.42
Vision-OPD (Yuan et al., 2026)	4B	89.53	59.41	81.75	80.12	74.51	70.17	75.92
Decomposed OPD (Yoon et al., 2026)	4B	89.80	58.70	82.60	79.50	73.80	67.82	75.37
CF-OPD (Ours)	4B	**90.58**	**59.88**	84.62	**81.50**	**77.78**	70.10	**77.41**
Qwen3.5-Based Models – 9B Scale								
Qwen3.5 (Qwen Team, 2026)	9B	86.39	52.31	85.12	80.88	71.13	67.31	73.86
GRPO (Shao et al., 2024)	9B	89.10	57.41	85.12	82.75	73.48	69.35	76.20
VA-OPD (Liu et al., 2026)	9B	86.91	62.84	86.12	82.75	71.17	70.20	76.67
V-Zero (Sun et al., 2026)	9B	89.53	57.04	85.50	**83.88**	**75.29**	**71.52**	77.13
Vision-OPD (Yuan et al., 2026)	9B	92.72	59.26	85.13	83.75	71.76	68.65	76.88
Decomposed OPD (Yoon et al., 2026)	9B	91.40	60.20	85.80	83.10	72.70	69.10	77.05
CF-OPD (Ours)	9B	**93.72**	**63.08**	**87.75**	83.50	72.21	70.46	**78.45**

4.3 CF-Bench-X: Responses to Semantic Edits

Figure 4 shows 77.6% Flip and 15.5% Blind for the 9B main-result model, outperforming the plotted Vision-OPD and GRPO models overall. Both measures matter: a model can change its answer after an edit without answering both images correctly. Flip therefore tests whether the response follows the intended semantic change, while Blind records unchanged predictions. Flip reaches 81.6% on Text, 86.1% on Color, 59.4% on Surface, and 74.5% on Shape. The Shape group improves over Vision-OPD by 16.4 points in Flip and reduces Blind from 25.5% to 12.7%. Color is strong for both methods, with CF-OPD ahead by 1.2 points in Flip and by 2.9 points in Blind. Flip is 47.8% on Object and 42.9% on State, evaluated on 23 and 14 pairs, respectively. GRPO also benefits from training on both images, highlighting the value of edited data. The same-input OPD comparison in Table 2 examines the additional benefit of using the pair within each target.

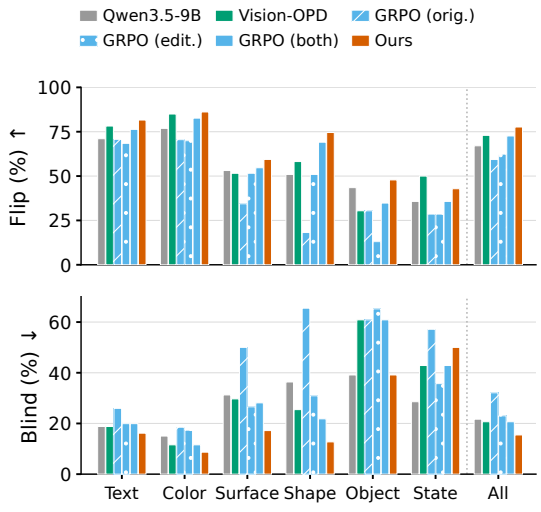


Figure 4: Flip (top) and Blind (bottom) by edit category on CF-Bench-X. Orange is ours.

Table 2: Training comparisons on Qwen3.5-9B; configurations are listed in Appendix A. Bold and underlining mark the best and second-best values in each metric column, including ties. Blind is lower-is-better; other metrics are higher-is-better.

Variant	V*	Zoom	HR4K	HR8K	MME-EN	MME-CN	Avg_6	Flip ↑	Blind ↓
Target variants									
CF-OPD (main result)	<u>93.72</u>	**63.08**	**87.75**	83.50	**72.21**	<u>70.46</u>	**78.45**	**77.6**	**15.5**
OPD target	91.10	<u>62.96</u>	86.25	**85.25**	71.13	69.43	77.69	71.1	21.0
CF-OPD ($\alpha = 0.25$)	92.67	**63.08**	<u>86.75</u>	<u>84.38</u>	<u>72.07</u>	70.24	<u>78.20</u>	<u>75.6</u>	<u>16.5</u>
no teacher mixture ($\lambda = 0$)	90.05	<u>62.96</u>	81.62	77.38	71.44	70.00	75.58	73.3	19.7
Baselines and other objectives									
Vision-OPD	92.72	59.26	85.13	83.75	71.76	68.65	76.88	72.9	20.7
GRPO, original images	87.43	55.27	78.12	75.50	70.87	**70.51**	72.95	59.3	32.1
GRPO, pairs	**94.76**	57.99	76.88	73.75	68.44	65.54	72.89	72.6	20.7
SFT, pairs	84.82	50.89	71.62	67.88	66.68	66.88	68.13	72.6	19.2

4.4 Ablations

Table 2 compares the 9B main model with target variants using the same training inputs and base checkpoint, alongside other training objectives. Appendix A provides input variants and hyperparameter sweeps, while Appendix B examines the gated target.

Paired versus independent targets. The *OPD target* baseline supervises the same images independently, while CF-OPD incorporates the partner crop into each target. CF-OPD reaches 78.45 Avg_6, 77.6 Flip, and 15.5 Blind, compared with 77.69, 71.1, and 21.0 for independent OPD. This improves Avg_6 by 0.76 points and Flip by 6.5 points, with Blind lower by 5.5 points. Recognition accuracy improves on five of six benchmarks, accompanied by stronger paired-answer performance.

Teacher mixture. The mixture comparison uses the same training inputs and fixes $\alpha = 0.25$. With $\lambda = 0$, CF-OPD records 75.58 Avg_6, 73.3 Flip, and 19.7 Blind. Setting $\lambda = 0.15$ raises Avg_6 to 78.20 and Flip to 75.6, while reducing Blind to 16.5. These gains of 2.62, 2.3, and 3.2 points support teacher mixing. The default model leads Table 2 on Avg_6, Flip, and Blind.

Comparison with Vision-OPD. Vision-OPD provides a reference at 76.88 Avg_6, 72.9 Flip, and 20.7 Blind. CF-OPD exceeds this baseline by 1.57 points in Avg_6 and 4.7 in Flip, with Blind lower by 5.2 points.

Other training objectives. For GRPO, using paired inputs increases Flip from 59.3 to 72.6 and reduces Blind from 32.1 to 20.7, while Avg_6 remains near 72.9. Paired SFT also obtains 72.6 Flip, alongside 68.13 Avg_6 and 19.2 Blind. Despite their identical Flip scores, paired GRPO and SFT differ by 4.76 points in Avg_6. These results motivate examining recognition accuracy together with both paired metrics when comparing training objectives.

5 Conclusion

Visual OPD transfers fine-detail recognition from a crop teacher to a full-image student. Semantic replacement supplements this supervision with teacher responses to controlled detail changes. We introduced CF-OPD to incorporate this response into supervision at student-generated prefixes, with paired crops contributing to target construction and valid edited images available for student training. Our editing and verification pipeline supplies reliable image-answer pairs, while CF-Bench-X measures correctness across both sides of a semantic change. CF-OPD achieves the highest Avg_6 among the compared Qwen3.5-based methods at both scales. On the same training inputs, CF-OPD improves over the independent OPD target (Table 2).

Limitations. Our evaluation covers still-image multiple-choice questions and a finite set of edit categories. Extending the pipeline to free-form answers and video would broaden its scope. Editor selection and generation throughput offer further opportunities to streamline offline data preparation.

Ethics Statement

The semantic edits are constructed for evaluating and improving visual recognition. Any release should identify edited images as synthetic and retain their edit metadata to distinguish them from the source images.

Reproducibility Statement

Section 4.1 specifies the training and evaluation settings. Appendix A reports the experimental configurations, and Appendix E provides the data-construction prompts. Code and dataset will be released upon acceptance.

AI Use Statement

We used AI-based assistants for coding support, drafting, and language editing. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final text, claims, and results.

References

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.

Yishuo Cai, Jiahui Liu, Yuanxin Liu, Haobo Deng, Linli Yao, Yuhao Zheng, Kun Ouyang, Zhimo Li, Ziyue Wang, Xu Sun, et al. Thinking without images: Internalizing visual manipulation with on-policy self-distillation. *arXiv preprint arXiv:2606.08719*, 2026.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

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.

Dongze Hao, Zhiwei Jin, Chen Chen, and Haonan Lu. Stabilizing on-policy distillation for mllm reasoning with global normalization. *arXiv preprint arXiv:2606.09091*, 2026.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, and Guohai Xu. Deepeyesv2: Toward agentic multimodal model. In *International Conference on Learning Representations*, volume 2026, pp. 114851–114872, 2026.

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

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *arXiv preprint arXiv:2607.21556*, 2026.

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.

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 2025. URL https://openai.com/index/introducing-gpt-5-2/. Published December 11, 2025.

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

Yunhang Qian, Jiaquan Yu, Jiawei Liu, Meng Wang, Hongwei Bran Li, and Xiaobin Hu. Medopd: 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, 2026. URL https://qwen.ai/blog?id=qwen3.5. Official model release and technical blog.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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.

Shravan Venkatraman, Omkar Thawakar, Ritesh Thawkar, Abdelrahman Shaker, and Rao Muhammad Anwer. Perception before supervision: Self-contained visual distillation from counterfactual blind spots. *arXiv preprint arXiv:2608.09931*, 2026.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

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

Shizhe Xiang, Ke An, Wenlong Yu, Yue Liu, Jian Luan, Pei Fu, and Qilong Wang. Teaching the way, not the answer: Privileged tutoring distillation for multimodal policy optimization. *arXiv preprint arXiv:2606.07000*, 2026.

Xi Xiao, Chen Liu, Chih-Ting Liao, Yunbei Zhang, Qizhen Lan, Yuxiang Wei, Lin Zhao, Janet Wang, Jianyang Gu, Muchao Ye, et al. Staying vigilant: Mitigating visual laziness via counterfactual visual alignment in mllms. *arXiv preprint arXiv:2606.26387*, 2026.

Feng Xiong, Leyan Xue, and Hongyu Lin. Correcting what you cannot see: Credit assignment for perception distillation in multimodal reasoners. *arXiv preprint arXiv:2607.28336*, 2026.

Leyan Xue, Feng Xiong, Mingjun Ma, and Changqing Zhang. Distill what the student can see: Fisher-projected on-policy distillation for vision-language models. *arXiv preprint arXiv:2608.01263*, 2026.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-opd: Confidence aware heterogeneous multi-teacher multimodal on-policy distillation. *arXiv preprint arXiv:2607.02592*, 2026.

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. *arXiv preprint arXiv:2606.00564*, 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.

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.

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? *arXiv preprint arXiv:2408.13257*, 2024.

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. *arXiv preprint arXiv:2508.11630*, 2025. doi: 10.48550/arXiv.2508.11630.

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. *arXiv preprint arXiv:2505.14362*, 2025.

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

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. Rp-opsd: Resolution-privileged on-policy self-distillation for multimodal large language models, 2026. URL https://arxiv.org/abs/2607.24447.

A Training Configurations and Hyperparameter Sweeps

Table 3 lists the runs behind Table 2 together with representative α and λ settings, separating input selection from the training objective. For the mixture sweeps, $q_t = (1-\lambda)q_t^* + \lambda\tilde{p}^+_{T,t}$, where λ is the teacher-mixture weight. The table records each run's training data and mixture settings. The learning rate and minibatch settings follow Section 4.1. Our evaluations on the six recognition benchmarks use Qwen3.5-397B for answer scoring. Flip and Blind are computed directly from option letters.

Parameter sensitivity. Recognition accuracy is stable across the nonzero-mixture α and λ sweeps with both training sides in the first block of Table 3: Avg_6 ranges from 77.65 to 78.45, a spread of 0.80 points. With $\lambda = 0.15$, varying α from 0.1 to 2 changes Avg_6 by only 0.25 points. These results suggest that competitive recognition performance does not require a narrowly tuned parameter choice within the tested range. Paired-answer metrics vary more: the same α sweep gives Flip between 75.6 and 77.6 and Blind between 15.5 and 16.6. At $\alpha = 0.25$, the nonzero-mixture λ sweep gives Flip between 71.9 and 76.5 and Blind between 16.5 and 20.0. We use $\alpha = 1$ and $\lambda = 0.15$ as the default.

Input and target comparisons. With $\alpha = 0.25$ and $\lambda = 0.15$, the main training setup reaches 78.20 Avg_6, 75.6 Flip, and 16.5 Blind. The paired-only run at these settings records 75.33 Avg_6, 77.0 Flip, and 16.3 Blind. For the same training inputs, the independent OPD target reaches 77.69 Avg_6, 71.1 Flip, and 21.0 Blind, compared with 78.45, 77.6, and 15.5 for the main CF-OPD model. This comparison uses the same training inputs for paired and independent targets.

Other objectives. GRPO with a multiple-choice reward on original images reaches 72.95 Avg_6, 59.3 Flip, and 32.1 Blind. With pairs, the corresponding scores are 72.89, 72.6, and 20.7. Supervised fine-tuning on original images reaches 70.96 Avg_6, 70.1 Flip, and 21.3 Blind; on edited images alone it reaches 58.21, 59.7, and 26.6; on pairs it reaches 68.13, 72.6, and 19.2.

B Gated Extension of CF-OPD

The default CF-OPD target is the fixed mixture in Eq. 4. We also evaluate a gated extension, denoted *CF-OPD (gated)*, which applies a smaller contrast adjustment at answer-related positions. Both model scales in Table 1 and the 9B analysis in Figure 4 use the default method. The experiments below evaluate the direction check and student input selection within this extension.

Gate and target. Let $m_t = 1$ when the input has a verified partner, both answer labels occur among the student's top-K candidates, and their combined mass averaged over $\tilde{p}^+_{T,t}$ and $\tilde{p}^-_{T,t}$ exceeds 0.5; otherwise $m_t = 0$. For each answer a, let $d_t(a)$ average δ_t over its available option-letter token forms, with and without a leading space. We compute

$$g_t = m_t \, \mathbb{1}\Big[\mathrm{TV}(\tilde{p}^+_{T,t}, \tilde{p}^-_{T,t}) \geq \tau \ \wedge \ d_t(y) > d_t(y')\Big]. \tag{9}$$

This checks relative answer preference under the two crops; it does not verify every candidate-token adjustment. With $Z_{T,t} = \sum_{v \in \mathcal{V}_t} p^+_{T,t}(v)$, the target is

$$Q^{\mathrm{g}}_t(v) = \begin{cases} (1 - \beta g_t) p^+_{T,t}(v) + \beta g_t Z_{T,t} q^*_t(v), & v \in \mathcal{V}_t, \\ 1 - Z_{T,t}, & v = \text{tail}. \end{cases} \tag{10}$$

Unlike the normalized top-K target of the default method, this extension preserves the crop teacher's tail mass. When $g_t = 0$, the target equals the ordinary crop teacher. Answer positions with $m_t = 1$ and $g_t = 0$ receive a loss multiplier of 0.9; other positions retain their original weight. A closed gate indicates a failed threshold or direction check, not proof that the teacher ignored the image.

Settings and results. The extension uses $\beta = 0.1$, $\tau = 0.2$, and $\kappa = 5$. The gated reference and its direction-check ablation use original student images, with edits as reference crops; the other rows change the input selection as indicated. Table 4 compares the 9B gated model with the direction inequality removed, and also reports the edited-only input variant. Removing the direction check

Table 3: Training configurations. CF and OPD denote paired and independent crop-teacher targets. Pairs means both pair members are used. The first block also includes original training examples; the second uses paired examples only. Bold and underlining mark best and second-best values, including ties; Blind is lower-is-better. Dashes indicate inapplicable settings.

Training input	Objective	α	λ	V*	Zoom	HR4K	HR8K	MME-EN	MME-CN	Avg_6	Flip	Blind
Main-setting comparisons and sweeps												
Pairs	CF (main)	1.0	0.15	93.72	63.08	**87.75**	83.50	72.21	70.46	**78.45**	**77.6**	15.5
Pairs	OPD	–	–	91.10	62.96	86.25	85.25	71.13	69.43	77.69	71.1	21.0
Pairs	CF	0.25	0	90.05	62.96	81.62	77.38	71.44	70.00	75.58	73.3	19.7
Pairs	CF (gated)	–	–	91.10	**64.62**	86.12	**85.38**	71.23	69.78	78.04	70.6	21.0
Pairs	CF	0.1	0.15	92.67	63.55	**87.75**	84.62	71.84	69.41	78.31	76.0	16.6
Pairs	CF	0.25	0.15	92.67	63.08	86.75	84.38	72.07	70.24	78.20	75.6	16.5
Pairs	CF	2.0	0.15	93.72	62.96	86.62	83.50	72.23	70.69	78.29	77.3	16.3
Pairs	CF	0.25	0.3	92.15	63.31	86.25	84.25	71.87	70.00	77.97	76.5	17.3
Pairs	CF	0.25	0.5	91.62	63.20	85.25	84.25	71.74	69.83	77.65	73.1	19.0
Pairs	CF	0.25	0.7	93.19	63.79	85.50	83.75	71.86	70.37	78.08	71.9	20.0
Pairs	Plain OPD	–	–	91.10	63.31	79.50	75.62	71.26	70.02	75.13	70.6	20.2
Paired-data comparisons and sweeps												
Pairs	CF	1.0	0.7	93.19	61.30	85.38	80.88	72.36	**71.59**	77.45	75.5	18.7
Pairs	CF	1.0	0.15	94.24	60.24	78.88	75.50	72.17	69.51	75.09	**77.6**	**15.1**
Pairs	CF	1.0	0.3	**95.29**	62.96	80.25	76.62	69.57	70.87	75.93	76.6	17.6
Pairs	CF	1.0	0.5	94.76	61.89	83.62	81.12	71.89	70.93	77.37	76.8	17.1
Pairs	CF	1.0	0	93.72	58.22	79.12	74.62	70.51	68.70	74.15	77.3	16.1
Pairs	CF	0.1	0.3	93.72	60.83	84.12	81.38	71.98	70.68	77.12	74.6	18.7
Pairs	CF	0.25	0.3	92.67	59.88	80.50	77.62	71.42	70.22	75.39	77.5	15.8
Pairs	CF	0.5	0.3	93.19	60.95	81.00	79.25	71.61	69.71	75.95	77.1	16.0
Pairs	CF	2.0	0.3	93.72	57.75	81.00	76.88	**72.68**	70.91	75.49	76.0	**15.1**
Pairs	CF	0.25	0.15	94.24	60.36	80.62	77.00	71.38	68.38	75.33	77.0	16.3
Pairs	OPD	–	–	87.43	62.72	80.75	80.12	71.18	69.55	75.29	73.8	19.7
Pairs	CF	1.0	0	93.19	57.87	79.12	74.50	71.42	68.85	74.16	77.1	17.3
Data–objective combinations												
Single side	CF	1.0	0.3	90.58	59.88	78.75	74.12	72.60	71.13	74.51	69.4	22.7
Original only	GRPO	–	–	87.43	55.27	78.12	75.50	70.87	70.51	72.95	59.3	32.1
Edited only	GRPO	–	–	84.29	46.63	79.38	73.50	61.96	52.93	66.45	62.4	23.0
Pairs	GRPO	–	–	94.76	57.99	76.88	73.75	68.44	65.54	72.89	72.6	20.7
Original only	SFT	–	–	87.96	56.33	74.75	72.12	67.83	66.76	70.96	70.1	21.3
Edited only	SFT	–	–	65.97	29.94	71.38	67.12	55.07	59.78	58.21	59.7	26.6
Pairs	SFT	–	–	84.82	50.89	71.62	67.88	66.68	66.88	68.13	72.6	19.2

lowers Avg_6 by 1.62 points and Flip by 1.4 points, and raises Blind by 1.2 points. Under the recipe of the main result, the gated target reaches 78.04 Avg_6 with 70.6 Flip and 21.0 Blind, 0.41 points below the default in Avg_6 and 7.0 points below it in Flip. The edited-only variant records 73.28 Avg_6, 72.4 Flip, and 21.3 Blind.

Table 4: Gated-extension experiments on Qwen3.5-9B. Bold and underlining mark the best and second-best values per metric, including ties; Blind is lower-is-better.

Variant	V*	Zoom	HR4K	HR8K	MME-EN	MME-CN	Avg_6	Flip ↑	Blind ↓
CF-OPD (gated)	89.53	63.79	**86.75**	**85.88**	**73.36**	**70.37**	**78.28**	70.8	22.2
main-result recipe	**91.10**	**64.62**	86.12	85.38	71.23	69.78	78.04	70.6	**21.0**
w/o direction check	85.86	63.79	86.25	83.88	70.66	69.55	76.66	69.4	23.4
edited inputs only	78.53	59.17	84.88	81.12	67.68	68.31	73.28	**72.4**	21.3

C Extended Related Work

Visual on-policy distillation. Vision-OPD (Yuan et al., 2026) introduced the paradigm: the student samples rollouts on the full image, and an EMA copy that sees the evidence crop supervises every sampled prefix. Follow-up work explores where the teacher's advantage should come from. RP-OPSD (Zhu et al., 2026) makes resolution the privilege, giving the teacher the native image and the student a downsampled copy. ViCuR (Tian et al., 2026) shows recoverable visual cues beat leaked answers as the privileged signal. S^2VOPD (Li et al., 2026) creates asymmetry by augmenting the student's view and distilling from the teacher's clean view. Imagine-OPD (Cai et al., 2026)

distills zoom-in tool trajectories into text so manipulation becomes imagination. H-OPD (Yin et al., 2026) routes each token's supervision to the most confident of several heterogeneous teachers. A second line refines the loss geometry. VA-OPD (Liu et al., 2026) concentrates the loss on tokens with high visual advantage between evidence-present and degraded views. VGS (Yoon et al., 2026) splits corrections into language-prior and visual-grounding parts and steers gradients to the latter. GNDPO (Hao et al., 2026) stabilizes training by batch-normalizing token-level advantages. FP-OPD (Xue et al., 2026) keeps only the correction components the student can express under a Fisher projection. PCD (Xiong et al., 2026) assigns credit separately to perception failures and reasoning failures. PTD-PO (Xiang et al., 2026) tutors failed rollouts on intermediate steps rather than final answers. These methods vary who the teacher is and how its signal is weighted. These approaches motivate examining how semantic replacement can supply a paired teacher comparison within the same training interface.

Counterfactual signals in visual distillation. A recent line estimates visual contribution by deleting evidence. VIGIL (Xiao et al., 2026) constructs a blinded state by ablating visual attention and penalizes answers that remain confident without vision, directly targeting visually lazy behavior. VCSD (Liang et al., 2026) erases the answer-relevant content to build a contrast image and turns the gap between the two teacher views into the asymmetry that drives self-distillation. V-Zero (Sun et al., 2026) removes answer labels entirely and keeps only trajectories whose predictions differ between positive and negative evidence views, so training data itself must demonstrate evidence use. Med-OPD (Qian et al., 2026) ports evidence-aware distillation to medical imaging, using a degraded modality as the counterfactual view for radiology question answering. Evidence-removal contrasts can also support visually attributed corrections and student-anchored target construction (Zhang et al., 2026). CVPD (Venkatraman et al., 2026) lets the model discover its own blind spots, then supervises them with a crop view as the positive teacher and a Gaussian-ghosted view as the negative teacher. These controls probe model responses to removing or degrading relevant information. Semantic replacement provides a comparison between two specified values of an answer-determining detail. CF-OPD verifies the resulting image-answer relation and uses the crop-teacher contrast to supervise student rollouts on both members of the pair.

D Target Properties

For two candidates $u, v \in \mathcal{V}_t$, the contrast-only target satisfies

$$\log \frac{q_t^*(u)}{q_t^*(v)} = \log \frac{p_{S,t}(u)}{p_{S,t}(v)} + \alpha\big(\delta_t(u) - \delta_t(v)\big). \tag{11}$$

Thus the contrast changes relative candidate preferences according to the teacher's estimated response. Residual editing artifacts and model errors can affect this estimate, and shared context does not imply exact cancellation of language priors. Equal contrast values preserve their relative odds in q_t^*; individual probabilities can still change through normalization, and the teacher mixture can change those odds further. Equation 11 describes the constructed target at one prefix, not a guaranteed separation between the student's final answers on two images.

E Data Construction Prompts

The data pipeline is driven by five prompts, executed in sequence by the annotation model and the image editor. Prompt 1 turns one annotated question into up to three edit plans; each plan names an edit type, a tight evidence box, and a target option that becomes the correct answer after the edit, with both images answered under the same question and option list. Plans may retain the source question or introduce a new one. Prompt 2 hardens a plan against mistargeting: it pins the object with an unambiguous description, lists the lookalike objects a careless editor might change instead, and writes a preserve clause for everything else. Prompt 3 illustrates the local-editing instruction, instantiated from the enriched plan and applied to the full-resolution evidence crop. The whole-frame route instead supplies the full image with the planned edit and preservation constraints. Prompts 4 and 5 verify the result from two independent angles: a blind reading that answers the question on both versions separately, and an informed audit that scores lighting, physics and factual coherence. A pair enters training only after passing both. We reproduce the prompts below, lightly condensed.

Prompt 1: Edit-Plan Generation

```
**Counterfactual Edit-Plan Generation**

# SYSTEM ROLE
You are building counterfactual training data for a fine-grained
visual perception model.  The image carries a red rectangle marking
the evidence region that decides the answer to the original question.

# INPUT
Original question, its options, the correct answer, and the evidence
box in 0--1000 normalised coordinates.

# TASK
Propose up to three counterfactual edit plans.  A plan repaints one
visible property of one local region so the correct answer changes
from one known value to another known value.  Plan 1 should keep the
original question and flip its answer to a different letter already
in the option list.  Plans 2 and 3 must use different edit types and
write their own four-option questions.  Fewer plans are fine; two good
plans beat three forced ones.

# ALLOWED EDIT TYPES
color, material, pattern, texture, text, shape, state, action_pose,
expression, orientation, object_identity, weather_time, lighting,
age_condition, transparency_clarity, clothing_attire, surface_marking.

# HARD CONSTRAINTS
Do not add or delete a person, animal, vehicle or accessory.  Do not
change the number of anything.  Do not move or resize objects.  Do not
require edits in two regions at once.  Do not redraw the whole scene
(weather_time and lighting excepted, only when the box covers the
sky or the lit area).  The edit must be a local repaint; everything
outside the box stays pixel-identical.

# BOX RULES
Boxes are [x1,y1,x2,y2] in 0--1000 coordinates, tight around the
edited object, each side above a minimum length, at most a bounded
fraction of the frame.  Boxes are re-validated in code.

# OUTPUT
JSON list of plans with edit_type, edit_instruction, box,
question, options, answer letters before and after the edit, and
use_original_question.
```

Prompt 2: Instruction Enrichment

```
**Anti-Mistargeting Enrichment**

# SYSTEM ROLE
You are preparing an instruction for an AI image editor.  The editor
is powerful but literal:  if the instruction does not pin down exactly
which object to change, it will change the wrong one, or repaint every
similar object in the scene.  A red rectangle marks the target.

# TASK
Given the planned edit, write four things.  (1) target_description:
identify the marked object so unambiguously that no other object
could be mistaken for it, including position in frame, current
appearance, and how it differs from same-kind objects; a noun phrase
of 15--40 words.  (2) target_short:  a 3--12 word appearance-only
phrase for close-up editing.  (3) lookalikes:  the other objects a
careless editor might change instead.  (4) preserve_clause:  one or
two sentences naming the lookalikes and their current appearance, and
stating that the rest of the scene, background, framing, viewpoint and
lighting stay unchanged.  Also classify edit_scope as single_object,
region, or whole_scene.

# OUTPUT
JSON with the five fields.
```

Prompt 3: Editing Instruction Template

```
**Precision Counterfactual Editing**

In this close-up, the subject is {target_short}.  {edit_instruction}.
{preserve_clause}.  Do not re-frame, do not zoom, do not change the
camera viewpoint, and do not add or remove any object.

The template is instantiated from Prompt 2's output and applied to the
full-resolution evidence crop; the edited crop is colour-matched on
a boundary ring and feathered back, so pixels outside the window stay
untouched.
```

Prompt 4: Blind Flip Verification

```
**Blind Flip Reading (four images)**

# INPUT
IMAGE 1 = version A, full scene, red rectangle on the questioned
region.  IMAGE 2 = that region enlarged.  IMAGE 3 = version B, full
scene, same rectangle.  IMAGE 4 = that region enlarged.

# TASK
Read fine detail off the enlarged views; use the full scenes only for
context.  Answer the question separately for version A and version
B, using only what you can see.  Then list every visible difference
between the two images, most obvious first, and count how many
separate places in the frame changed.

# RULES
One option letter per image, judged independently; use "?" if an image
does not decide it.  Report only what you can see; do not speculate
about why the images differ.

# OUTPUT
JSON with answer_image1, answer_image2, confidences, differences,
n_regions_changed.
```

Prompt 5: Logic Audit

```
**Informed Audit:  Lighting / Physics / Factual Logic**

# SYSTEM ROLE
You are auditing an AI image edit before it enters a training set.  Be
strict:  a bad pair teaches a perception model the wrong thing, so
when in doubt, mark it down.

# INPUT
The original image, the edited image, the intended edit instruction,
and the intended region in 0--1000 coordinates.

# CRITERIA
1.  EXECUTED: the intended change was carried out clearly and
completely.
2.  LOCALITY: nothing outside the intended region changed, including
background, other objects, faces, framing, colour cast, and text
elsewhere.
3.  LIGHTING (0--3):  shadow direction and length agree with the
scene; highlights sit where the light source is; colour temperature
matches; changed light sources update the light they cast; no halo,
seam, or grain mismatch.
4.  PHYSICS (0--3):  the object still rests on, attaches to, and is
occluded by the same things; perspective and scale agree; materials
behave like themselves; state changes carry their consequences (a lamp
turned off no longer lights nearby surfaces).
5.  FACTUAL (0--3):  text is real and correctly spelled; objects keep
real-world-valid combinations; anatomy is correct; the object is one
identifiable thing; the scene remains possible.

# OUTPUT
JSON with executed, locality_ok, outside_changes, lighting_score,
physics_score, factual_score, verdict.
```

F Qualitative Examples Across Edit Directions

Our editing taxonomy contains seventeen types across four broad families: appearance (color, material, pattern, texture, transparency/clarity), written content (text, surface marking), identity (object identity), and state or configuration (shape, state, action/pose, expression, orientation, weather/time, lighting, age/condition, clothing/attire). Figures 5–7 show verified pairs for six representative edit types in the dataset-release format. These examples use crop-and-paste editing, with the red box highlighting the intended evidence region. Each pair carries two known option-letter answers to the same question. All shown pairs passed the pixel gate, blind answer check, and logic audit. The examples illustrate local color and text changes as well as replacements involving materials, patterns, object identity, and state.

original



edited (counterfactual)



Before edit - original image

Q: What is the primary color of the car shown in the image?
A. white B. red C. black D. blue
Answer: C. black

Edit instruction

Change the car's paint from black to blue

After edit - counterfactual image

Q: (same question, asked on the edited image)
Answer: D. blue

original



edited (counterfactual)



Before edit - original image

Q: What word is written in large yellow letters on the front banner of the stall?
A. Rojak B. Borneo C. Mix D. Fruit
Answer: B. Borneo

Edit instruction

Change the large yellow text on the banner from 'Borneo' to 'Mix'

After edit - counterfactual image

Q: (same question, asked on the edited image)
Answer: C. Mix

Figure 5: Verified counterfactual pairs: color edit (top) and text edit (bottom).

original



edited (counterfactual)



Before edit - original image

Q: What material is the base of the statue made of?
A. metal B. wood C. concrete D. stone
Answer: D. stone

Edit instruction

Change the stone base of the statue to look like weathered wood with visible grain and knots

After edit - counterfactual image

Q: (same question, asked on the edited image)
Answer: B. wood

original



edited (counterfactual)



Before edit - original image

Q: What pattern is on the dress worn by the woman walking away on the left?
A. solid B. polka dot C. striped D. floral
Answer: C. striped

Edit instruction

Change the striped pattern on the woman's dress to a solid black color

After edit - counterfactual image

Q: (same question, asked on the edited image)
Answer: A. solid

Figure 6: Verified counterfactual pairs: material edit (top) and pattern edit (bottom).

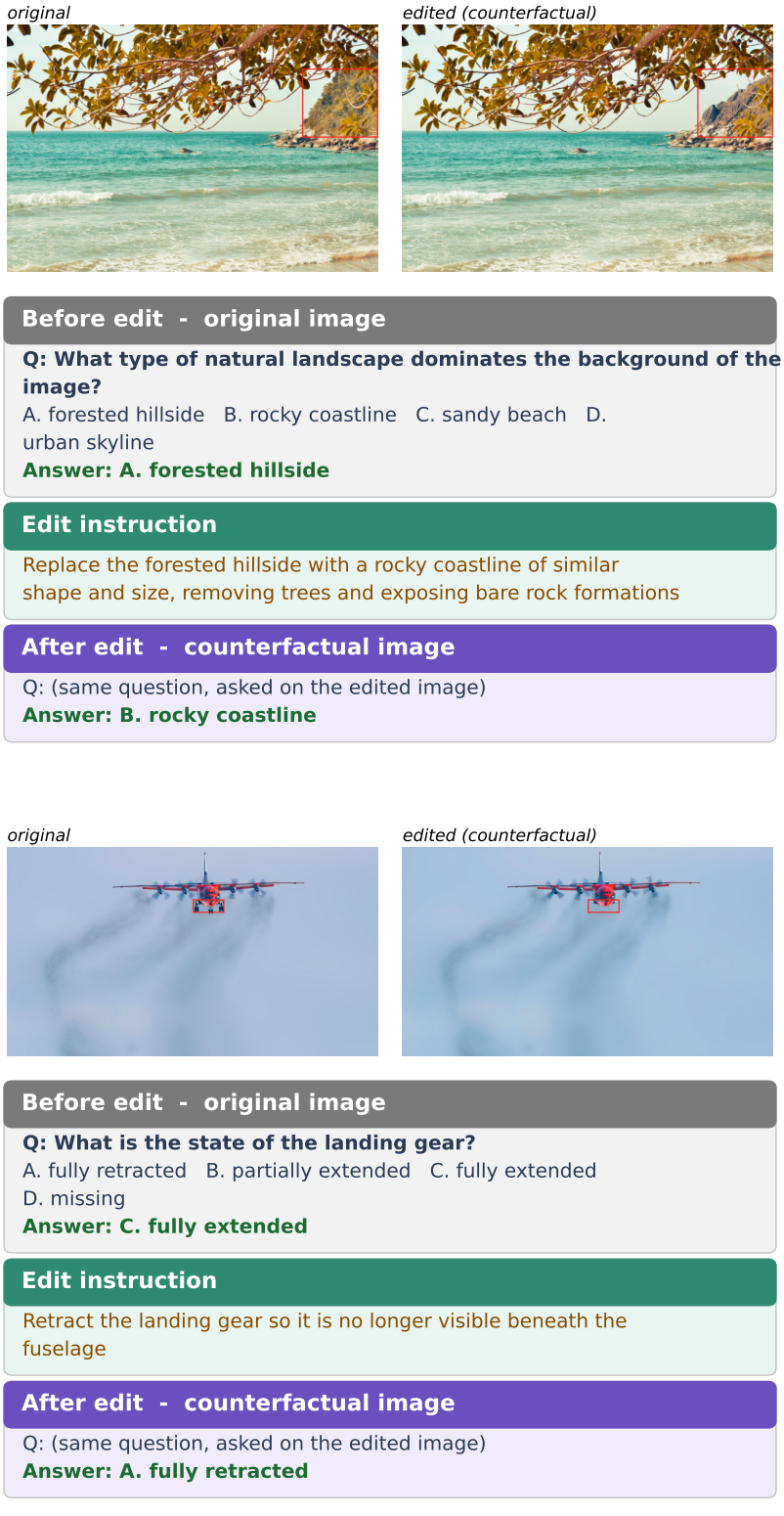


Figure 7: Verified counterfactual pairs: object identity edit (top) and state edit (bottom).