DISTILL THE VISUAL EVIDENCE, NOT JUST THE ANSWER: CROSS-WORLD ON-POLICY DISTILLATION FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A central goal of vision-language model (VLM) distillation is to transfer both the teacher's language capabilities and its visual understanding. However, existing methods primarily supervise the student's output, leaving visual understanding implicit. Our analysis reveals that a student can match the teacher's answer without relying on the same visual evidence, raising the question: how can we ensure the student responds to the visual information that actually determines the answer? To this end, we propose **Cross-World On-Policy Distillation (CW-OPD)**, which explicitly supervises the student's response to changes in visual evidence. For each example, CW-OPD constructs two visual worlds that share the question and scene context but differ in answer-critical evidence, yielding different answers. We perform on-policy distillation in both worlds and distill the teacher's cross-world belief transition, encouraging the student to match not only *what* the teacher predicts but also *why* its prediction changes with the evidence. A gradient analysis shows that this term is invariant to errors shared by both worlds and supplies a corrective signal invisible to endpoint matching alone. In this way, CW-OPD makes reliance on the relevant visual evidence an explicit distillation target rather than an implicit consequence of output matching. To diagnose whether a model truly grounds its answers in visual evidence, we introduce CWBench, which measures cross-world consistency via Cross-World Pair Accuracy (CWPA). Experiments on Qwen3.5-4B show that CW-OPD outperforms the strongest baseline by **1.2** points on average, and the 4B student exceeds DeepSeek-V4.1 (552B) by **22.4** CWPA points on CWBench. Code is included in the supplementary material.

1 INTRODUCTION

On-policy distillation (OPD) has emerged as an effective paradigm for distilling large language models (LLMs), where the teacher supervises responses generated by the student itself (Agarwal et al., 2024; Gu et al., 2024). This on-policy design aligns the teacher and student state distributions, substantially reducing the noise and distribution mismatch introduced during distillation (Agarwal et al., 2024; Ko et al., 2024). However, vision-language models (VLMs) jointly rely on textual and visual information (Liu et al., 2023; Dai et al., 2023), making it challenging for OPD to transfer the teacher's visual capabilities through supervision in the textual generation space (Yoon et al., 2026).

Recent efforts to improve multimodal distillation can be broadly divided into two directions. The first revisits the distillation objective itself, aiming to isolate or reweight the visual component so that the student receives more effective gradients for visual learning. For example, Visual Gradient Steering (VGS) (Yoon et al., 2026) observes that the language-alignment and visual-perception gradients occupy nearly orthogonal subspaces, and dynamically steers the optimization toward the visual subspace to prioritize visual learning. The second direction introduces privileged visual conditions that provide stronger supervision than the student's standard input. Vision-OPD (Yuan et al., 2026) uses a crop of the region of interest (ROI) as a privileged teacher view for full-image distillation. However, both directions improve *how effectively* visual supervision is transferred, without constraining *which* visual evidence the student's answer depends on. As shown in Fig. 1(a), the teacher and student can produce the same correct answer while attending to markedly different visual regions, with the student often assigning less cross-modal attention to the answer-critical evidence.

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.

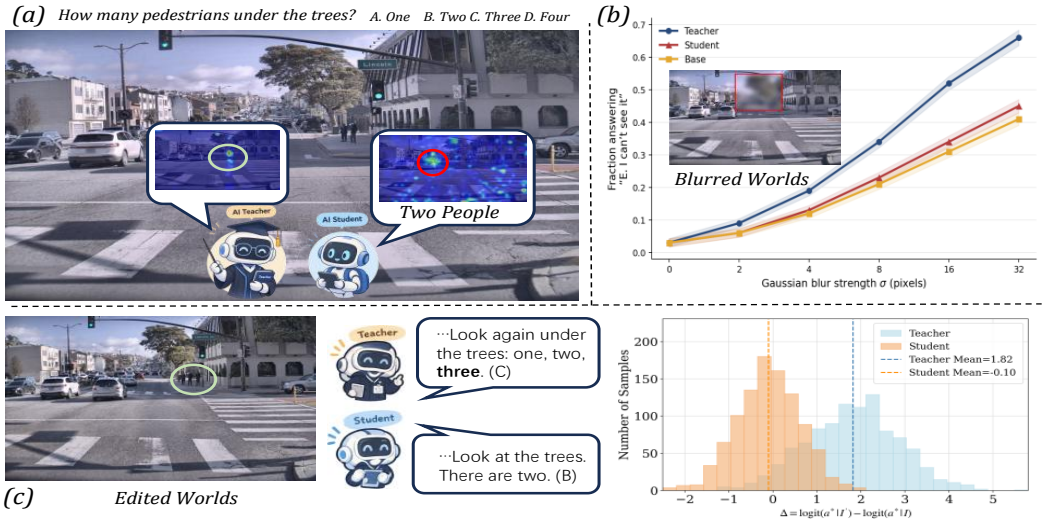

Figure 1: Motivating observations on the training dataset. (a) Teacher and student exhibit different cross-modal attention despite producing the same answer. (b) As the target region is progressively blurred, the student's sensitivity to the option "E. I can't see it" barely differs from the base model. (c) Under semantically edited worlds, the student fails to distinguish between them, whereas the teacher significantly increases the logit of the correct answer a^* in the edited world (Welch's $t = 47.6$, $p < 10^{-4}$, Cohen's $d = 2.13$), confirming the significance of this effect. Student: Qwen3.5-4B. Results average the Qwen3.5-9B teacher (OPD) and the 4B EMA teacher (OPSD).

We further quantify this evidence–output misalignment from two perspectives. First, output sensitivity (Fig. 1b): masking the answer-critical region causes a substantially larger change in the teacher's answer logits than in the student's, indicating that the student's prediction is less dependent on the relevant visual evidence. Second, we construct semantically contrasted visual worlds in which the answer-critical evidence is altered while the question and broader scene context remain unchanged (Fig. 1c). Under this controlled intervention, the teacher exhibits a consistent shift in its token probabilities, whereas the student's response is significantly less aligned with the teacher's belief transition. These observations reveal a common limitation of existing methods: matching output distributions in a single visual world does not ensure that the student grounds its answer in the answer-critical evidence.

To address these limitations, we propose **Cross-World On-Policy Distillation (CW-OPD)**, which reformulates visual knowledge transfer as a cross-world distillation problem. We first construct CWBench, a dataset of dual-world pairs (I, q) and (I', q) that share the question and broader scene context but differ in answer-critical evidence, such that the two worlds support different answers. We then perform on-policy distillation in both worlds and further distill the teacher's cross-world belief transition, requiring the student to reproduce not only what the teacher predicts at each world but also why its prediction changes when the visual evidence is altered. This design suppresses world-invariant shortcuts while directly supervising the visual response relevant to the task. Extensive experiments demonstrate that CW-OPD improves visual grounding and downstream VLM performance across diverse benchmarks. Our contributions are summarized as follows:

- We identify an overlooked limitation of on-policy distillation for VLMs: matching the teacher's output distribution does not necessarily ensure that the student relies on the answer-critical visual evidence, leaving room for linguistic and other evidence shortcuts.
- We propose **Cross-World On-Policy Distillation (CW-OPD)**, which constructs semantically symmetric cross-world training pairs and combines bidirectional on-policy distillation with cross-world belief transition matching to supervise both world-specific predictions and their responses to visual changes. We release **CWBench**, a dataset based on this pipeline.
- Experiments across diverse benchmarks show that CW-OPD improves visual grounding and downstream performance, surpassing the strongest baseline and even frontier models on cross-world consistency.

2 Related Work

2.1 On-Policy Distillation

On-policy distillation (OPD) (Agarwal et al., 2024) addresses the distribution mismatch of teacher-forced distillation by evaluating the teacher on student-generated trajectories, providing token-level supervision on visited states. Follow-up works explore different divergences and sampling strategies, e.g., MiniLLM (Gu et al., 2024), GKD, and DistiLLM (Agarwal et al., 2024; Ko et al., 2024), focusing on output-level knowledge transfer. For VLMs, however, predictions may rely on different visual evidence, which standard OPD does not explicitly constrain.

2.2 Multimodal OPD: Privileged Views and Visual Objectives

Recent multimodal OPD methods improve visual knowledge transfer in two directions. One introduces privileged visual views. Vision-OPD instantiates a crop-conditioned teacher and a full-image student from the same model, minimizing token-level divergence along the student's on-policy rollouts to internalize the benefit of visual zooming without inference-time tools (Yuan et al., 2026). RP-OPSD instead supervises a low-resolution student with a teacher conditioned on the original resolution, so that the resolution gap serves as privileged information requiring only image–question pairs (Zhu et al., 2026a). In both cases, the teacher's privileged input creates an information gap relative to the student's deployed view.

Another line of work modifies the distillation objective itself. VGS decomposes the OPD loss via Bayes' rule into a language-prior term and a visual-grounding term, and steers the update toward the visual subspace since the two gradients are nearly orthogonal (Yoon et al., 2026). FP-OPD projects the teacher–student log-probability gap onto the student's local visual tangent space under the Fisher metric, distilling only locally realizable corrections (Xue et al., 2026). VA-OPD defines a per-token visual advantage from the teacher's log-probability change when visual detail is destroyed, and reweights rollouts and token groups by it (Liu et al., 2026). VAD reconstructs the target rather than its weight: it projects the teacher correction onto a signed proxy of the visual evidence direction obtained by contrasting evidence-present and evidence-removed views (Zhang et al., 2026).

These methods strengthen, filter, or reweight visual supervision, but mainly constrain predictions under individual visual conditions. In contrast, CW-OPD constructs two semantic worlds with different answer-critical evidence via visual intervention, and distills both predictions and the teacher's cross-world belief transition, explicitly supervising the student's response to visual evidence changes.

3 Method

3.1 Revisiting On-Policy Distillation

On-Policy Distillation. On-policy distillation (OPD) (Agarwal et al., 2024) provides token-level supervision by matching the student's next-token distribution with that of a fixed teacher along trajectories generated by the student itself. Given a visual-language input (I, q), where I denotes the image and q denotes the question, the student policy π_θ first generates an on-policy trajectory $\tau \sim \pi_\theta(\cdot \mid I, q)$. For each generated prefix $y_{<t}$, the student and teacher next-token distributions and the OPD objective are

$$p^I_{S,t} = \pi_\theta(\cdot \mid I, q, y_{<t}), \qquad p^I_{T,t} = \pi_{\bar{\theta}}(\cdot \mid I, q, y_{<t}),$$
$$\mathcal{L}_{\mathrm{OPD}} = \frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\left(p^I_{S,t} \,\|\, p^I_{T,t}\right), \tag{1}$$

where $\pi_{\bar{\theta}}$ denotes a frozen teacher policy (or an exponential-moving-average teacher). Unlike conventional offline distillation, the supervision is conditioned on the student's own generation trajectory, allowing the teacher to provide dense feedback at the states that the student actually visits.

Linguistic Shortcut. Although OPD provides token-level supervision, its objective is defined on the marginal predictive distribution rather than the visual evidence underlying the prediction. Let

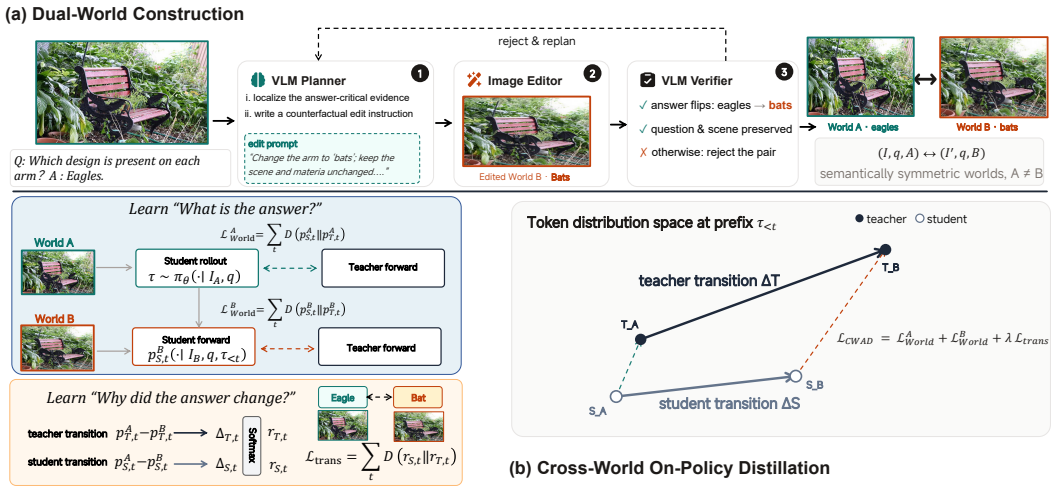

Figure 2: **(a) Dual-world construction:** a VLM planner writes a counterfactual edit instruction, an image editor produces the edited image, and a VLM verifier accepts the pair only if the answer flips while the question and scene are preserved. **(b) CW-OPD:** the student rolls out once in World A and all passes reuse its prefix $\tau_{<t}$; bidirectional OPD anchors both worlds ($\mathcal{L}^A_{World} + \mathcal{L}^B_{World}$), and the transition loss $\mathcal{L}_{trans}$ distills how the teacher's belief changes with the evidence.

Z denote the latent visual evidence extracted from image I for question q. The student's predictive distribution can be written as

$$p_S(y \mid I, q) = \sum_z p_S(y \mid z, q)\, p_S(z \mid I, q), \tag{2}$$

where $p_S(z \mid I, q)$ characterizes the evidence relied upon by the student. Substituting Eq. 2 into the reverse KL used by OPD gives

$$D_{\mathrm{KL}}\left(p_S(\cdot \mid I, q), \|, p_T(\cdot \mid I, q)\right) = \mathbb{E}_{y \sim p_S(\cdot \mid I, q)}\left[\log \frac{\sum_z p_S(y \mid z, q), p_S(z \mid I, q)}{\sum_z p_T(y \mid z, q), p_T(z \mid I, q)}\right]. \tag{3}$$

Thus, OPD only requires the student to match the teacher's marginal prediction after aggregating over possible evidence z; it does not separately constrain which visual evidence gives rise to that prediction. Consequently, a correct output can be achieved through different evidence distributions $p_S(z \mid I, q)$, including reliance on non-critical visual cues or linguistic priors.

3.2 Cross-World On-Policy Distillation

Dual-World Formulation. Motivated by the evidence ambiguity of single-world supervision, we construct semantically symmetric visual world pairs via a plan–edit–verify pipeline (Figure 2a). Given an original world (I, q, A), a VLM planner localizes the answer-critical evidence and writes a counterfactual edit instruction; an image editor then produces I', yielding

$$(I, q, A) \quad \longleftrightarrow \quad (I', q, B), \qquad A \neq B, \tag{4}$$

where B is an alternative answer compatible with the question and scene. A VLM verifier accepts the pair only if the answer flips ($A \rightarrow B$) and the question and non-critical scene context are preserved; otherwise the planner rewrites the instruction. The accepted worlds share the question and overall scene, but exchange answer-critical evidence, so each is a valid visual input inducing a different answer. Details are in Sec. 4.2 and Appendix A.

World Matching and Cross-World Transition. Given a dual-world pair, the student should reproduce the teacher's behavior in both worlds while preserving their semantic contrast. The student rolls out once in world A, and the resulting prefix is shared across both worlds (Algorithm 1 in Appendix B). For the student-generated trajectories, let $p^A_{S,t}$ and $p^A_{T,t}$ denote the student and teacher logits in world A, and $p^B_{S,t}$ and $p^B_{T,t}$ those in world B. We first apply OPD to each world:

$$\mathcal{L}_{\text{world}} = \frac{1}{T}\sum_{t=1}^{T}\left[D_{\mathrm{KL}}\left(p^A_{S,t} \| p^A_{T,t}\right) + D_{\mathrm{KL}}\left(p^B_{S,t} \| p^B_{T,t}\right)\right], \tag{5}$$

which anchors the two semantic endpoints. Endpoint matching alone, however, does not supervise how the prediction changes when the visual evidence is altered. Following the comparative distillation formulation of Xu et al. (2025), we therefore distill the cross-world belief transition as the KL divergence between the softmax-normalized per-token log-likelihood ratios $\Delta_{S,t} = \log p^A_{S,t} - \log p^B_{S,t}$ and $\Delta_{T,t} = \log p^A_{T,t} - \log p^B_{T,t}$:

$$\mathcal{L}_{\text{trans}} = \frac{1}{T} \sum_{t=1}^{T} D_{\text{KL}} \left(\text{softmax}(\Delta_{S,t}) \parallel \text{softmax}(\Delta_{T,t}) \right), \tag{6}$$

where $\text{softmax}(\Delta_{S,t}) \propto p^A_{S,t}/p^B_{S,t}$ denotes the relative odds between the two worlds induced by the visual intervention. Here the comparison operates on a semantically counterfactual world pair under on-policy distributions, rather than on independent samples. It captures both the direction and magnitude of the belief change, encouraging the student to reproduce why the teacher's prediction changes when the visual evidence is altered, not merely what it predicts at each endpoint.

Overall Objective. The final CW-OPD objective combines the two complementary constraints:

$$\mathcal{L}_{\text{CW-OPD}} = \mathcal{L}_{\text{world}} + \lambda \mathcal{L}_{\text{trans}}, \tag{7}$$

where λ balances per-world prediction matching against cross-world transition distillation. Figure 2(b) visualizes the computation of both losses in the token-distribution space at a shared prefix $\tau_{<t}$: the endpoint KL divergences and the cross-world log-ratio vectors Δ_S, Δ_T are computed from the four distributions $p^A_{S,t}, p^A_{T,t}, p^B_{S,t}, p^B_{T,t}$ and aggregated into $\mathcal{L}_{\text{CW-OPD}}$ by Eq. equation 7.

3.3 When and Why Does Transition Supervision Help?

What does the transition term add to endpoint supervision? We answer by decomposing the gradients of the two losses at one scored position. Let ℓ_E and ℓ_T be the per-position endpoint and transition losses inside the sums of equation 5 and equation 6, with $\mathcal{E}, \mathcal{T}$, and $\mathcal{J}_\lambda = \mathcal{E} + \lambda\mathcal{T}$ their averages over scored positions. All gradients below condition on the sampled prefix $\tau_{<t}$, treating the discrete on-policy rollout as fixed, i.e., they are gradients of the surrogate objective along one update. Splitting each world's Jacobian into its common-mode and image-dependent parts, $\bar{J} = (J_A + J_B)/2$ and $J_\Delta = J_A - J_B$ with $J_w = \partial z^w_S / \partial\theta$, and writing $\varepsilon_w = F(p^w_S)(\log p^w_S - \log p^w_T)$ for the reverse-KL residual in world w, direct differentiation gives (derivation in Appendix C)

$$\nabla_\theta \ell_E = \bar{J}^\top(\varepsilon_A + \varepsilon_B) + \tfrac{1}{2} J_\Delta^\top(\varepsilon_A - \varepsilon_B), \qquad \nabla_\theta \ell_T = J_\Delta^\top F(r_S)(d_S - d_T). \tag{8}$$

Eq. equation 8 separates each gradient into a common-mode part, acting through $\bar{J}$, and an image-dependent part, acting through J_Δ. For the endpoint loss, both parts are present. Its common-mode part is driven by the sum $\varepsilon_A + \varepsilon_B$, which collects teacher errors made identically in both worlds, such as a language prior at the shared prefix, the shared hint, or a calibration offset. The transition loss has no common-mode part at all: its gradient lives entirely in the image-dependent subspace of $J_\Delta^\top$. Proposition 1 states the two properties formally; proofs are in Appendix C.

Proposition 1 (Common-mode rejection)**.** *(i) Adding the same vector b to both teacher logits (or both student logits) leaves ℓ_T unchanged, while ℓ_E changes in general. (ii) The transition gradient vanishes along any direction that moves both worlds identically ($J_A\delta = J_B\delta \Rightarrow \nabla_\theta \ell_T \cdot \delta = 0$).*

Proposition 1 clarifies how the transition term interacts with shortcuts. Part (i) removes *additive* logit components shared by the two worlds, such as a common calibration offset or a language-prior shift at the shared prefix, from the transition target instead of reproducing them at both endpoints. Part (ii) gives zero transition gradient to parameter directions that move both worlds identically, so the signal reaches only parameters whose responses differ across worlds; this motivates keeping the vision encoder trainable, so that such parameters exist. At an endpoint-stationary point $\nabla\mathcal{E}(\theta_{\text{sc}}) = 0$ with $\nabla\mathcal{T}(\theta_{\text{sc}}) \neq 0$, Eq. equation 8 then implies that a step along the transition gradient decreases $\mathcal{T}$ to first order, $\mathcal{T}(\theta_{\text{sc}} - \eta\nabla\mathcal{T}) = \mathcal{T}(\theta_{\text{sc}}) - \eta\|\nabla\mathcal{T}\|_2^2 + O(\eta^2)$, while $\mathcal{E}$ changes only at $O(\eta^2)$: transition supervision provides a first-order descent direction for the cross-world mismatch that is invisible to endpoint matching at a stationary point. The transition term thus supervises *how* the student's prediction responds to the evidence change, not only *what* it predicts. Whether this alignment improves answers is an empirical question. We address it with the structured perturbation experiments in Appendix C.

Table 1: Main results. All methods are evaluated at the original image resolution. Avg. is the unweighted mean over the seven standard benchmarks. Bold denotes the best result. On CWBench we report CWPA (Eq. 9) and CWFR (Eq. 10).

Model	V*Bench	HR-Bench 4K	RealWorldQA	MMVP	HallusionBench	SEED-Bench	OK-VQA	Avg.	CWPA ↑	CWFR ↓
Qwen3.5-0.8B										
Base	63.87	47.09	58.56	67.21	62.36	70.44	51.50	60.15	37.71	50.00
SFT	67.21	48.53	58.42	70.46	61.97	71.22	55.80	61.94	37.90	49.72
GRPO	64.60	44.50	58.67	70.39	63.83	70.91	53.20	60.87	37.65	50.11
OPSD	71.20	69.00	**66.27**	67.00	**64.84**	74.32	**62.15**	67.83	41.94	48.33
OPD	70.68	70.25	65.10	70.67	63.95	70.69	61.40	67.53	39.19	48.53
Vision-OPD	70.83	50.47	63.24	71.12	63.28	71.78	55.00	63.67	38.43	49.31
VAD	74.42	70.13	63.66	70.72	62.48	74.01	57.38	67.54	42.71	49.38
VA-OPD	73.93	70.26	64.82	69.45	62.63	72.11	59.82	67.57	39.73	50.32
FP-OPD	71.20	70.07	65.40	70.72	64.13	71.51	60.50	67.65	39.02	48.87
VGS	72.42	69.25	65.12	69.89	63.73	70.81	58.80	67.15	38.04	49.58
RP-OPSD	**75.39**	70.13	64.71	70.68	61.12	73.22	54.04	67.04	41.73	48.32
CW-OPD	72.77	**70.75**	**66.27**	**71.32**	64.13	**75.56**	60.88	**68.81**	**48.28**	**43.79**
Qwen3.5-4B										
Base	81.15	83.00	74.38	77.67	70.42	79.25	75.17	77.29	47.64	46.13
SFT	83.50	83.75	73.30	78.67	68.03	77.29	77.42	77.42	48.57	45.38
GRPO	83.60	80.88	73.50	81.67	71.30	79.70	76.90	78.22	48.09	45.91
OPSD	84.82	84.63	78.43	79.82	73.67	78.52	78.75	79.81	60.61	33.50
OPD	82.70	83.25	76.84	79.17	68.17	80.31	76.56	78.14	49.53	44.02
Vision-OPD	84.10	83.90	77.20	80.00	69.90	80.55	76.80	78.92	50.33	43.47
VAD	84.54	83.67	78.32	81.13	68.43	**80.79**	**78.93**	79.40	54.32	42.57
VA-OPD	85.69	84.32	76.29	78.67	72.34	79.96	78.33	79.37	54.27	40.96
FP-OPD	83.67	83.50	77.80	79.33	69.50	80.60	77.30	78.81	49.05	44.36
VGS	83.40	83.30	77.10	79.00	69.20	80.40	76.90	78.47	48.81	44.71
RP-OPSD	85.75	**85.50**	77.20	79.50	71.85	78.90	76.80	79.36	56.16	37.94
CW-OPD	**86.39**	85.12	**80.65**	**82.00**	**74.58**	80.10	77.96	**80.97**	**65.82**	**29.30**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Training Settings. CW-OPD is instantiated on top of on-policy self-distillation (OPSD), where the frozen teacher is an EMA copy of the student conditioned on a short answer hint generated by the base model as a teacher-side privileged input; an instantiation with a standard larger teacher (OPD) is provided in Appendix D. We train on *CW-Bench*, our dedicated benchmark for cross-world distillation, which provides 2,365 training dual-world pairs and a 594-pair evaluation set; the data are drawn from VLM-CapCurriculum Perception (Wu et al., 2026), ZwZ-RL-VQA (Wei et al., 2026), Visual7W (Zhu et al., 2016), and the Vision-OPD dataset (Yuan et al., 2026). For each example, CW-OPD constructs a dual-world pair $(I, q, A) \leftrightarrow (I', q, B)$ through the plan–edit–verify pipeline. All student parameters, including the vision encoder, are trainable. We optimize $\mathcal{L}_{\mathrm{CW-OPD}} = \mathcal{L}_{\mathrm{world}} + \lambda \mathcal{L}_{\mathrm{trans}}$ with $\lambda = 1.0$; across a wide range of λ, the transition loss converges at different rates but reaches similar final performance (Appendix F). We sample one response per pair at temperature 2.0. All models are trained for one epoch with a learning rate of 2×10^{-6} on 4 NVIDIA RTX PRO 6000D GPUs (96 GB); the global batch size is 64 for Qwen3.5-0.8B and 48 for Qwen3.5-4B, determined by GPU memory constraints. More detailed experimental setups are provided in the Appendix E.

Benchmarks. We evaluate on standard benchmarks across three categories—visual perception and grounding (V*Bench (Wu & Xie, 2024), HR-Bench 4K (Wang et al., 2024), RealWorldQA xAI (2024), MMVP (Tong et al., 2024)), hallucination diagnosis (POPE (Li et al., 2023b), Hallusion-Bench (Guan et al., 2024)), and external-knowledge reasoning (SEED-Bench (Li et al., 2023a), OK-VQA (Marino et al., 2019))—together with the CWBench evaluation set. All models are evaluated at the original image resolution; model responses are parsed with rule-based matching where applicable, and otherwise scored by an LLM judge (Qwen3.5-9B). Detailed dataset statistics, task formulations, and evaluation protocols are provided in Appendix G. On CW-Bench we report CWPA and CWFR (Eqs. 9 and 10); see §4.2 for the benchmark construction, metric definitions.

Baselines. Our main comparison includes the Base model and standard post-training methods, including SFT, GRPO (Shao et al., 2024), OPD, and OPSD (Zhao et al., 2026). We further compare with representative multimodal distillation methods that either construct privileged views, such as RP-OPSD (Zhu et al., 2026b), Vision-OPD (Yuan et al., 2026), VAD Zhang et al. (2026) and VA-OPD Liu et al. (2026), or modify the distillation objective, such as FP-OPD (Xue et al., 2026) and VGS (Yoon et al., 2026). All methods are trained on the training split of CWBench under identical training settings for fair comparison, with details provided in Appendix E.

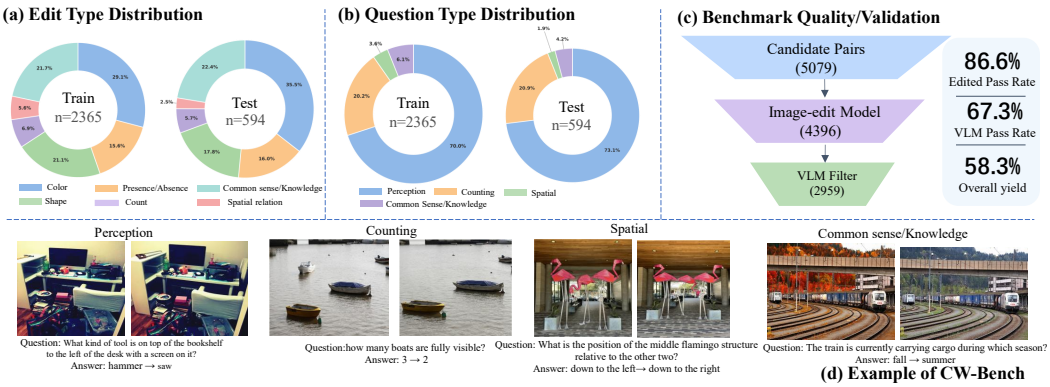

Figure 3: Statistics, validation, and examples of the proposed CW-Bench. (a) Edit type distribution of the answer-critical visual interventions in the train (n=2,365) and test (n=594) splits. (b) Question type distribution of the paired queries in both splits. (c) The filtering pipeline for constructing dual-world pairs: from 5,079 candidate pairs, 4,396 pairs pass the image-edit validation (86.6%), and 2959 pairs remain after VLM filtering (67.3% pass rate, 58.3% overall yield). (d) Examples of dual-world pairs: for each example, the original and edited images are shown with the shared question; only the answer-critical evidence differs, flipping the answer between the two worlds.

Main Results. CW-OPD consistently improves overall performance across both model scales. On Qwen3.5-0.8B, CW-OPD achieves an average score of **68.81**, surpassing the strongest OPD-family baseline (FP-OPD, 67.65) by 1.16 points and the strongest OPSD-family baseline (OPSD, 67.83) by 0.98 points. The gain is more pronounced on Qwen3.5-4B, where CW-OPD reaches **80.97**, improving over Vision-OPD (78.92) by 2.05 points and OPSD (79.81) by 1.16 points. These results demonstrate that the benefits of cross-world distillation generalize across model scales, with consistent improvements over existing on-policy distillation approaches. Beyond standard benchmarks, CW-OPD yields substantially larger gains on our cross-world diagnostic: CWPA increases from 37.71 to **48.28** on 0.8B and from 47.64 to **65.82** on 4B, while CWFR drops from 50.00 to **43.79** and from 46.13 to **29.30**, respectively. The 18.2-point CWPA gain on 4B confirms that the improvement directly reflects the intended effect of CW-OPD: the student becomes more sensitive and consistent to answer-critical changes in visual evidence across paired worlds.

4.2 The CWPA Benchmark

Construction and evaluation metrics. Dual-world pairs are constructed via a plan–edit–verify pipeline (Appendix A) and a three-stage filtering cascade (Fig. 3c): from 5,079 candidates, 86.6% pass image-edit validation and 67.3% survive VLM filtering, yielding 2,959 verified pairs (58.3% yield) in which the answer flips while the question and scene context are preserved. Let $\hat{y}_W$ and a_W denote the model's and gold answers in world $W \in \{A, B\}$. We report the per-world accuracies A Acc $= \frac{1}{N}\sum_i \mathbb{1}[\hat{y}_A^{(i)}{=}a_A^{(i)}]$ (likewise B Acc), and the *Cross-World Pair Accuracy*

$$\mathrm{CWPA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\hat{y}_A^{(i)}{=}a_A^{(i)} \wedge \hat{y}_B^{(i)}{=}a_B^{(i)}\right], \tag{9}$$

which requires answering *both* worlds correctly and thus measures whether correctness survives the evidence change. To isolate inconsistency from raw accuracy across models at different accuracy regimes, we further report the *flip rate*, the fraction of pairs whose correctness differs between the two worlds:

$$\mathrm{CWFR} = \mathrm{A\ Acc} + \mathrm{B\ Acc} - 2\,\mathrm{CWPA}. \tag{10}$$

Diversity. CW-Bench covers a wide range of interventions and reasoning skills (Fig. 3a,b). Answer-critical edits span color, presence/absence, shape, object count, spatial relations, and commonsense-knowledge changes, with no single edit type dominating either split. Paired questions concentrate on fine-grained perception (70.0%/73.1% in train/test) and counting (20.2%/20.9%), with spatial and commonsense reasoning making up the remainder—requiring models to ground answers in visual evidence rather than linguistic priors.

Table 2: Cross-world consistency of frontier models and our distilled students on CWBench. We report CWPA (Eq. 9) and CWFR (Eq. 10); higher CWPA and CWPA-Q and lower CWFR indicate more consistent evidence-grounded answering.

Model	Size	A Acc	B Acc	CWPA↑	CWPA-Q↑	CWFR↓
Kimi-K3	2.8T	70.03	72.90	50.00	62.14	42.93
DeepSeek-V4.1	552B	54.38	79.63	43.43	64.32	47.15
Qwen3.8	27B	72.05	79.29	57.58	69.93	36.18
Qwen3.5(Base)	4B	61.95	79.46	47.64	59.86	46.13
OPSD	4B	74.75	79.97	60.61	66.45	33.50
CW-OPD	4B	**78.11**	**82.83**	**65.82**	**73.60**	**29.30**

Table 3: Ablation of CW-OPD supervision components. **DW-OPD** applies bidirectional per-world OPD to both worlds but *omits* the transition loss($\mathcal{L}^A_{\mathrm{OPD}} + \mathcal{L}^B_{\mathrm{OPD}}$), isolating the contribution of cross-world transition supervision. We report the CWBench metrics CWPA (Eq. 9) and CWFR (Eq. 10), together with Avg., the unweighted mean over the seven standard VQA benchmarks.

Method	$\mathcal{L}^A_{\mathrm{OPD}}$	$\mathcal{L}^B_{\mathrm{OPD}}$	$\mathcal{L}_{\mathrm{trans}}$	A Acc.	B Acc.	CWPA ↑	CWFR ↓	Avg.
base	–	–	–	61.95	79.46	47.64	46.13	77.29
OPD	✓	–	–	60.92	82.16	49.53	44.02	78.14
OPSD	✓	–	–	74.75	79.97	60.61	33.50	79.81
DW-OPD	✓	✓	–	72.69	81.33	59.12	35.78	79.43
CW-OPD	✓	✓	✓	**78.11**	**82.83**	**65.82**	**29.30**	**80.97**

Analysis. Table 2 admits two observations. First, within the same model family, CW-OPD substantially strengthens the base model's visual evidence understanding (CWPA 47.6→65.8) and, more importantly, surpasses OPSD with the same privileged hint (60.6 CWPA / 33.5 CWFR), showing that the gain comes from cross-world transition supervision rather than from the hint itself. Second, commercial frontier models that adopt on-policy distillation in post-training—as documented in their technical reports—exhibit exactly the deficiency diagnosed in Fig. 1: their per-world accuracy is strong, yet their correctness does not survive evidence changes, leaving CWPA far below what their per-world accuracy promises and CWFR as high as 47.2. This suggests that the shortcut is a property of the OPD paradigm rather than of model scale or of small-scale distillation. Notably, a 4B student trained with CW-OPD surpasses all of them in both CWPA (65.8) and CWFR (29.3) at an order-of-magnitude smaller scale. To verify that this consistency is selective rather than indiscriminate, we additionally report CWPA-Q, which replaces the question with one unrelated to the edited region and checks whether the answer follows the question instead of the visual edit. CW-OPD attains the highest CWPA-Q (73.6), clearly above OPSD (66.4) and all frontier models, indicating that the student answers what the question asks and ignores the irrelevant intervention—precisely the selective evidence response that per-world supervision alone fails to enforce.

4.3 Ablation Study

Ablation of supervision components. Table 3 disentangles the contribution of the three supervision signals. Per-world OPD alone improves single-world accuracy but leaves cross-world consistency nearly unchanged, confirming that output matching in one world does not constrain how the prediction responds to evidence changes. Simply extending OPD to both worlds (DW-OPD) anchors the two semantic endpoints but can even enlarge the gap between them, as the student grows more confident in each world without learning to *transition* between them—the correctness sets of the two worlds drift apart. Only when the transition loss is added does the student align its belief change with the teacher's, yielding simultaneous gains in per-world accuracy, CWPA, and downstream benchmarks. This validates our design: per-world distillation and cross-world transition supervision are complementary and jointly necessary—the former fixes *where* the prediction sits in each world, while the latter fixes *how* it moves when the evidence changes.

Ablations on transition supervision. Figure 4 examines what makes cross-world supervision effective from two complementary angles. First, the alignment objective must capture the *full* belief transition rather than a projection of it: cosine alignment, which preserves only the direction of the transition, and magnitude alignment, which preserves only its norm, both leave a large answer-

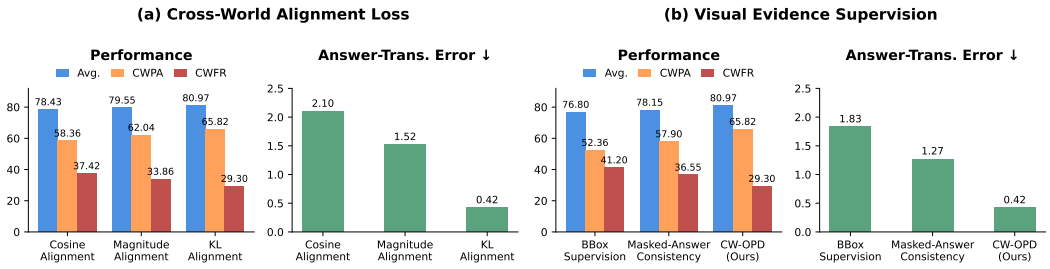

Figure 4: Ablation studies on (a) the design of the cross-world transition loss and (b) the visual evidence supervision strategy, evaluated on the 4B student. Each panel reports the Avg. score over seven standard VQA benchmarks, CWPA and CWFR on CWBench, and the *Answer-Transition Error* (ATE): the average absolute gap between the student and the teacher in the log-likelihood change of the gold answer token between the two worlds, which is the same quantity diagnosed in Fig. 1(c). (a) Cosine, magnitude, and KL variants of the transition loss. (b) BBox supervision, masked-answer consistency, and the full CW-OPD.

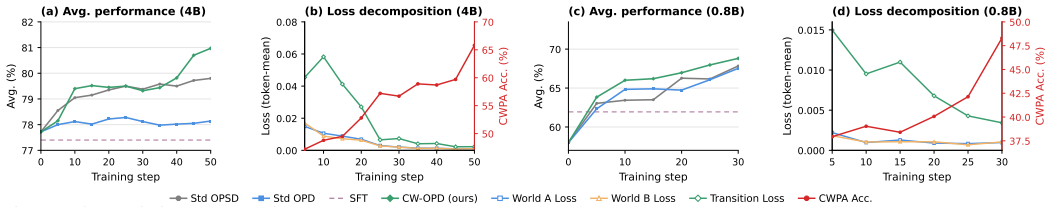

Figure 5: Training dynamics on the Student 4B (A) and Student 0.8B (B) models. (a, c) CWPA over training steps for SFT, OPD, OPSD, and CW-OPD. (b, d) Decomposition of the CW-OPD training objective into World A loss, World B loss, and cross-world transition loss.

transition error and degrade CWBench performance; KL alignment, which matches the teacher's transition as a distribution, is the only variant that simultaneously reduces ATE, CWFR, and improves downstream accuracy. Second, the supervision must be anchored to *answer-relevant* evidence: BBox supervision improves spatial grounding but leaves the answer transition largely unlearned, and masked-answer consistency enforces the answer without constraining the evidence behind it; only CW-OPD, which supervises how the answer-relevant belief changes between worlds, attains the best performance across all metrics. Together these results indicate that the effectiveness of CW-OPD stems not from stronger supervision per se, but from supervising the *right transition*—the teacher's response to the evidence change on the answer—in the *right space*—the full distributional response rather than any single projection of it.

Training dynamics. Figure 5 tracks CWPA throughout training. Standard post-training methods (SFT, OPD, OPSD) plateau near or below the pretrained level, confirming that conventional objectives do not improve—and may even erode—cross-world consistency. In contrast, CWPA rises monotonically under CW-OPD and surpasses all baselines within a few steps, with the gain correlating closely with the decay of the transition loss: as $\mathcal{L}_{\text{trans}}$ decreases, the student progressively aligns its belief shift with the teacher's, while the per-world losses remain stable anchors. This indicates that CWPA is actively optimized by the cross-world objective rather than improved as a by-product of better single-world fitting.

5 Conclusion

We identify a limitation of on-policy distillation for vision-language models: matching the teacher's output distribution within a single visual world does not ensure that the student relies on the answer-critical evidence, leaving room for world-invariant shortcuts. To make evidence reliance an explicit supervision target, we propose Cross-World On-Policy Distillation (CW-OPD), which constructs semantically symmetric dual-world pairs and supervises both per-world predictions and the cross-world belief transition. Together with CWBench, which measures cross-world consistency through CWPA and CWFR, our experiments show that CW-OPD improves visual grounding and downstream performance across model scales, and that frontier models trained with on-policy distillation remain inconsistent under evidence changes. A promising next step is to extend cross-world supervision to reinforcement learning, which may also reduce the reliance on image editing to construct dual-world pairs and thus cover evidence types that are hard to alter faithfully.

AI USE STATEMENT

During the preparation of this work, the authors used AI-assisted tools (Large Language Models) in the following capacities: (1) constructing and curating portions of the CWBench dataset, including generating and rephrasing candidate edit instructions and questions; (2) assisting with code development, debugging, and documentation; and (3) polishing and improving the clarity of the manuscript's writing. All AI-generated content was reviewed and verified by the authors, who take full responsibility for the accuracy, originality, and integrity of the paper and its artifacts.

REFERENCES

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895, 2024.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3195–3204, 2019.

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

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Weiyun Wang, Lewei Ding, Mengchen Zeng, Xinghang Zhou, Lijun Shen, Yadong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *arXiv preprint arXiv:2408.15556*, 2024.

Long Wei, Lianyu He, Jiaxu Lan, Luoxi Dong, Ying Cai, Shuo Li, Hao Zhu, Wenxuan Wang, Ling Kong, Yiding Wang, Zehuan Zhang, and Weilin Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Jiayu Wu, Han Chen, Hao Tu, Xiaotao Tang, Feng Shi, Huan Liu, Hao Lu, Cihang Xie, and Yibo Zhou. From seeing to thinking: Decoupling perception and reasoning improves post-training of vision-language models. *arXiv preprint arXiv:2605.20177*, 2026.

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.

xAI. Realworldqa. `https://x.ai/news/realworldqa`, 2024.

Alex Tianyi Xu, Alex Wilf, Paul Pu Liang, Alexander Obolenskiy, Daniel Fried, and Louis-Philippe Morency. Comparative knowledge distillation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 7690–7699, 2025.

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.

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, Shijian Wang, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. VAD: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026.

Shijing Zhao, Zhen Xie, Ming Liu, Jian Huang, Guansong Pang, Feng Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. RP-OPSD: Resolution-privileged on-policy self-distillation for multimodal large language models. *arXiv preprint arXiv:2607.24447*, 2026a.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. Rp-opsd: Resolution-privileged on-policy self-distillation for multimodal large language models. *arXiv preprint arXiv:2607.24447*, 2026b.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

A CW-BENCH

This appendix details the construction of **CW-Bench**, the dual-world dataset used for Cross-World On-Policy Distillation. CW-Bench is built from a general VQA pool covering five sources: `zwz_rl_vqa_mcq`, `vision_opd_original_no_box_dual_resolution`, `vlm_capcurriculum_perception`, `A-OKVQA`, and `visual7w(cauldron,llava_format)`. We keep samples whose `task_family` is `mcq` and whose ground-truth answer appears among the options. The resulting pool contains 5,079.

A.1 Counterfactual Edit Instruction Generation

For each sample, the planner receives the image, question, options, and the ground-truth answer. It selects a distractor option that can be made true by a single local edit, and then writes an English edit instruction for the image editor. The selection follows four hard criteria:

1. The edit touches only one object or one attribute; everything else must stay pixel-familiar.
2. After the edit, the chosen option is unambiguously true and the ground-truth answer is unambiguously false for this question. Reject any option that would leave two plausible answers.
3. Reject options that require:
   - tiny or invisible details that cannot be clearly seen in the image;
   - a different camera viewpoint or major scene restructuring.
4. The instruction must (a) name what is currently in the image and where it is, (b) say what to change it into while keeping the same pose/size/position, (c) pin the material and style to the original, and (d) explicitly list what to keep unchanged (other parts, background, framing, lighting).

Text editing (signs, labels, logos, jerseys, scoreboards, banners, license plates, book covers, etc.) is now allowed, as Qwen-Image-3.0 can handle text modifications well. Our prompt is:

Counterfactual Edit Instruction Prompt

```
You are preparing a counterfactual image for a visual-editing API (
    Qwen-Image-3.0).
The image shown is the ORIGINAL. Below is a 4-option VQA question
    about it and its ground truth (GT).

Question: {question}
Options: {options}
GT: {gt}. {gt_text}

Task: Pick ONE distractor option (not the GT) that can become TRUE in
    the image via a SINGLE LOCAL edit,
and write the English editing instruction for it.

Hard criteria for the chosen option:
1. The edit touches ONE object or ONE attribute only; everything else
    must stay pixel-familiar.
2. After the edit the chosen option is UNAMBIGUOUSLY true and the GT
    UNAMBIGUOUSLY false for this
  question. Reject any option that would leave two plausible answers.
3. Reject options that require:
  - Tiny or invisible details that cannot be clearly seen in the image
      .
  - A different camera viewpoint or major scene restructuring.
4. The instruction must (a) name what is CURRENTLY in the image and
    where it is, (b) say what to
  change it into, keeping the same pose/size/position, (c) pin the
      material and style to the original,
  (d) explicitly list what to keep unchanged (other parts, background,
      framing, lighting).

NOTE: Text editing (signs, labels, logos, jerseys, scoreboards,
    banners, license plates, book covers, etc.)
is NOW ALLOWED. Qwen-Image-3.0 can handle text modifications well.

Example (different image, for format only):
{few_shot}

Answer with ONE JSON object and nothing else:
```

```
{"chosen_option": "<letter or null>", "chosen_text": "<option text>",
    "edit_scope": "<the object/attribute touched>", "edit_instruction":
     "<English instruction, 2-4 sentences>", "editability": <1-5>}
```

If no distractor satisfies these conditions, the planner outputs `chosen_option = null`. The output is a strict JSON object with fields `chosen_option`, `chosen_text`, `edit_scope`, `edit_instruction`, and `editability` (1–5). We run the planner with a Qwen-based VLM backend (Qwen3.8-27B) using `max_tokens=1024` and `temperature=0.4`, with automatic retries for rate limits. The pipeline supports sharding by `idx` and idempotent resumption.

A.2 Image Editing

Given the generated instructions, we produce counterfactual images with the Qwen-Image-3.0 API. The editor outputs a square image of 1024×1024. Image editing succeeds for 4,396 out of 5,079 attempted samples (86.6%). The total editing time is 16.94 hours with two workers, at a cost of approximately 643 RMB. The prompt is:

Counterfactual Edit Instruction Prompt

```
You are preparing a counterfactual image for a visual-editing API (
    Qwen-Image-3.0).
The image shown is the ORIGINAL. Below is a 4-option VQA question
    about it and its ground truth (GT).

Question: {question}
Options: {options}
GT: {gt}. {gt_text}

Task: Pick ONE distractor option (not the GT) that can become TRUE in
    the image via a SINGLE LOCAL edit,
and write the English editing instruction for it.

Hard criteria for the chosen option:
1. The edit touches ONE object or ONE attribute only; everything else
    must stay pixel-familiar.
2. After the edit the chosen option is UNAMBIGUOUSLY true and the GT
    UNAMBIGUOUSLY false for this
   question. Reject any option that would leave two plausible answers.
3. Reject options that require:
   - Tiny or invisible details that cannot be clearly seen in the image
       .
   - A different camera viewpoint or major scene restructuring.
4. The instruction must (a) name what is CURRENTLY in the image and
    where it is, (b) say what to
   change it into, keeping the same pose/size/position, (c) pin the
       material and style to the original,
   (d) explicitly list what to keep unchanged (other parts, background,
        framing, lighting).

NOTE: Text editing (signs, labels, logos, jerseys, scoreboards,
    banners, license plates, book covers, etc.)
is NOW ALLOWED. Qwen-Image-3.0 can handle text modifications well.

Example (different image, for format only):
{few_shot}

Answer with ONE JSON object and nothing else:
{"chosen_option": "<letter or null>", "chosen_text": "<option text>",
    "edit_scope": "<the object/attribute touched>", "edit_instruction":
     "<English instruction, 2-4 sentences>", "editability": <1-5>}
```

A.3 Quality Control and Verification

A VLM verifier examines each candidate pair $(I, q, A) \leftrightarrow (I', q, B)$ and accepts it only if (i) the answer flips from A to B and (ii) the question and the non-critical scene context are preserved. Rejected pairs are sent back to the planner for instruction rewriting. After quality control, we obtain 2,959 verified dual-world pairs for training and analysis. The prompt is:

Faithfulness Score Prompt

```
You are given TWO images of the same scene. The FIRST image is the
    ORIGINAL. The SECOND image is the EDITED version produced by an
    image-editing model.

Editing instruction given to the model: "{edit_instruction}"

What the edit targets: "{edit_scope}"

Score how faithfully the SECOND image executes the instruction, on a
    continuous scale from 0.0 to 1.0:
- 1.0: the described change is fully and unambiguously realized in the
     edited image, AND the rest of the image (background, other objects
    , composition, lighting) stays consistent with the original.
- 0.75: the target change is clearly realized with minor imperfections
    ; rest well preserved.
- 0.5: the edit attempted the change but result is ambiguous, partial,
     or the wrong object changed.
- 0.25: barely related to the instruction, or large unintended changes
     elsewhere.
- 0.0: no change at all, or the image is corrupted/unrelated.

Answer with ONE JSON object and nothing else:
{"score": <float 0.0-1.0>, "reason": "<one short sentence>"}
```

A.4 Edit and Question Type Classification

We automatically classify the 2,959 verified edits using Qwen3.5-9B in a pure text classification setting, with zero parsing errors. The prompt is:

Edit type and question type classification.

Edit Type / Question Type Classification Prompt

```
Classify the following image editing task and question into the
    categories below.

**Edit instruction:** {edit_instruction}
**Edit scope:** {edit_scope}
**Chosen answer:** {chosen_text}
**Question:** {question}

Output JSON with exactly these fields:
- "edit_type": one of ["count", "color", "shape", "presence/absence",
    "spatial relation", "common-sense/knowledge"]
- "question_type": one of ["perception", "counting", "spatial
    reasoning/knowledge", "others"]

Edit type definitions:
- count: editing the number/quantity of objects (e.g., "change 3 dogs
    to 5 dogs")
- color: editing the color of something (e.g., "change red to blue")
```

Algorithm 1 One CW-OPD update (per training pair)

Require: pair (I^A, I^B, q); student π_θ; teacher π_T (frozen or EMA); optional teacher hint h shared by both worlds; weight λ

1: sample $\tau \sim \pi_\theta(\cdot \mid I^A, q)$ at the rollout temperature; let T be its number of valid response positions (first response token through EOS; prompt and padding excluded)
2: **for** $t = 1, \ldots, T$ with the same tokens $\tau_{<t}$ **do**
3: $p^w_{S,t} = \pi_\theta(\cdot \mid I^w, q, \tau_{<t})$ for $w \in \{A, B\}$ (with gradient)
4: $p^w_{T,t} = \pi_T(\cdot \mid I^w, q, h, \tau_{<t})$ for $w \in \{A, B\}$ (no gradient)
5: $\ell_{E,t} = D_{\mathrm{KL}}(p^A_{S,t} \| p^A_{T,t}) + D_{\mathrm{KL}}(p^B_{S,t} \| p^B_{T,t})$
6: $\Delta_{M,t} = \log p^A_{M,t} - \log p^B_{M,t}$, $r_{M,t} = \mathrm{softmax}(\Delta_{M,t})$ for $M \in \{S, T\}$, $\ell_{T,t} = D_{\mathrm{KL}}(r_{S,t} \| r_{T,t})$
7: **end for**
8: $\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \big(\ell_{E,t} + \lambda\, \ell_{T,t}\big)$; update θ with the optimizer of Table 6
9: **note:** $\lambda = 0$ gives DW-OPD with identical rollouts, scoring, masks and reductions

```
- shape: editing the shape/form/design of something (e.g., "change
    eagle shape to bat shape")
- presence/absence: adding or removing an object entirely (e.g., "add
    a cat", "remove the person")
- spatial relation: editing position/orientation/posture (e.g., "
    change lying to standing")
- common-sense/knowledge: editing based on world knowledge (e.g., "
    change to African setting")

Question type definitions:
- perception: asking about visual attributes (color, shape, texture,
    design)
- counting: asking about number/quantity
- spatial reasoning/knowledge: asking about position, action, or
    requiring world knowledge
- others: anything else

Output only the JSON, no explanation.
```

Summary. CW-Bench provides 2,959 verified dual-world pairs with diverse edit types and question categories. Each pair shares the same question and preserves the overall scene, while exchanging answer-critical evidence so that the two worlds induce different answers.

B Algorithm

Algorithm 1 summarizes one CW-OPD update. The student rolls out once in World A, and the resulting prefix $\tau_{<t}$ is reused for both worlds, so the two next-token distributions differ only in the visual evidence. At each position we compute the per-world student and teacher distributions, form the per-world KL term $\ell_{E,t}$, and then distill the teacher's cross-world belief transition through the centered log-ratio response distributions $r_{S,t}$ and $r_{T,t}$. The final loss combines the two terms with weight λ; setting $\lambda = 0$ recovers DW-OPD under identical rollouts, scoring, masks, and reductions.

C Proofs and Mechanism Verification

C.1 Proof of Proposition 1

Recall that $d_M = z^A_M - z^B_M$ and $r_M = \mathrm{softmax}(d_M)$ for $M \in \{S, T\}$.

Part (i). Adding b to both teacher logits changes the teacher transition logits to $(z^A_T + b) - (z^B_T + b) = z^A_T - z^B_T = d_T$; hence r_T and the target of ℓ_T are unchanged, and so is ℓ_T itself. The

Table 4: CWPA and Avg. under structured teacher perturbations (4B student, Qwen3.5-9B teacher; values averaged over 3 seeds). Clean conditions reproduce Table 3.

	Teacher perturbation				
Method	Clean	Shared β=0.5	Shared β=1.0	Opp. β=0.5	Opp. β=1.0
CWPA ↑					
DW-OPD (λ=0)	59.12	$53.0^{\pm 1.2}$	$47.0^{\pm 1.4}$	$57.5^{\pm 1.1}$	$56.0^{\pm 1.2}$
CW-OPD (λ=1)	65.82	$64.6^{\pm 0.9}$	$64.0^{\pm 1.0}$	$60.0^{\pm 0.8}$	$55.2^{\pm 1.3}$
Avg. over seven benchmarks ↑					
DW-OPD (λ=0)	79.43	$77.8^{\pm 0.3}$	$76.1^{\pm 0.3}$	$79.0^{\pm 0.2}$	$78.6^{\pm 0.1}$
CW-OPD (λ=1)	80.97	$80.8^{\pm 0.2}$	$80.5^{\pm 0.2}$	$79.9^{\pm 0.2}$	$78.9^{\pm 0.3}$

same argument applies to the student logits. In contrast, the endpoint target in world w becomes $\mathrm{softmax}(z_T^w + b)$, which differs from $\mathrm{softmax}(z_T^w)$ unless $b \propto \mathbf{1}$, so ℓ_E changes in general.

Part (ii). From Eq. equation 8, $\nabla_\theta \ell_T = J_\Delta^\top F(r_S)(d_S - d_T)$. If $J_A\delta = J_B\delta$, then $J_\Delta\delta = J_A\delta - J_B\delta = 0$, and therefore $\nabla_\theta \ell_T \cdot \delta = \big(F(r_S)(d_S - d_T)\big)^\top J_\Delta \delta = 0$. □

C.2 Derivation of Eq. equation 8

For a fixed token position, let d_S denote the student transition logits and $r_S = \mathrm{softmax}(d_S)$ with Jacobian $F(r_S) = \mathrm{diag}(r_S) - r_S r_S^\top$. Differentiating $D_{\mathrm{KL}}(r_S \| r_T) = \sum_i r_{S,i}(\log r_{S,i} - \log r_{T,i})$ with respect to d_S gives

$$\nabla_{d_S} D_{\mathrm{KL}}(r_S \| r_T) = F(r_S)(d_S - d_T), \tag{11}$$

because $\nabla_{d_S} \log r_S = I - \mathbf{1} r_S^\top$ and $F(r_S)\mathbf{1} = 0$. Note that this gradient vanishes whenever $d_S - d_T \propto \mathbf{1}$, since the softmax removes constant shifts; non-vanishing therefore requires $d_S - d_T \notin \mathrm{span}\{\mathbf{1}\}$. Since $d_S = z_S^A - z_S^B$, $\partial d_S / \partial \theta = J_\Delta$, which yields the second line of Eq. equation 8.

For the endpoint loss, write the per-position loss as $D_{\mathrm{KL}}(p\|q)$ with $p = \mathrm{softmax}(z)$ and a fixed target q. Using $\partial p_i / \partial z_j = F_{ij}(p)$ and $\partial \log p_i / \partial z_j = \delta_{ij} - p_j$,

$$\frac{\partial D_{\mathrm{KL}}(p\|q)}{\partial z_j} = \sum_i F_{ij}(p)\,(\log p_i - \log q_i) + \sum_i p_i(\delta_{ij} - p_j) = \big[F(p)(\log p - \log q)\big]_j,$$

since the second sum equals $p_j - p_j = 0$. Applying this in each world with the residual $\varepsilon_w = F(p_S^w)(\log p_S^w - \log p_T^w)$ gives $\nabla_\theta \ell_E = J_A^\top \varepsilon_A + J_B^\top \varepsilon_B$; substituting $J_A = \bar{J} + \frac{1}{2} J_\Delta$ and $J_B = \bar{J} - \frac{1}{2} J_\Delta$ yields the first line of Eq. equation 8.

Remark (approximate stationarity). Section 3.3 states the shortcut argument at an exact stationary point $\nabla \mathcal{E}(\theta_{\mathrm{sc}}) = 0$. If only $\|\nabla \mathcal{E}(\theta_{\mathrm{sc}})\|_2 \le \varepsilon$, the same Taylor expansion adds a term of order $\eta\varepsilon$ to the endpoint loss, and $\mathcal{T}$ still decreases to first order whenever $\lambda \|\nabla \mathcal{T}\|_2 > \varepsilon$; the mechanism is therefore robust to small violations of exact stationarity.

C.3 Structured teacher perturbations (target-invariance sanity check)

Proposition 1(i) implies an invariance property of the transition *target*: adding the same logit vector b to both worlds leaves d_T, and hence r_T, unchanged, whereas opposite perturbations offset the teacher contrast by $2b$. We use structured teacher-logit perturbations as a sanity check of this target-level property on the Qwen3.5-4B student with the frozen Qwen3.5-9B teacher. At every scored position we add $b = \beta \hat{b}$ to the teacher logits, where $\hat{b}$ is a fixed random direction normalized to unit norm and scaled by the per-token logit standard deviation; *shared* applies the same b in both worlds, *opposite* applies $+b$ in world A and $-b$ in world B. We report CWPA on the same split, prompt, and parser as Table 3, averaged over 3 seeds.

The observed robustness ordering is consistent with this invariance: shared perturbations, which do not alter the transition target, affect DW-OPD much more strongly than CW-OPD, whereas opposite perturbations, which corrupt the teacher contrast, affect CW-OPD much more strongly. Note

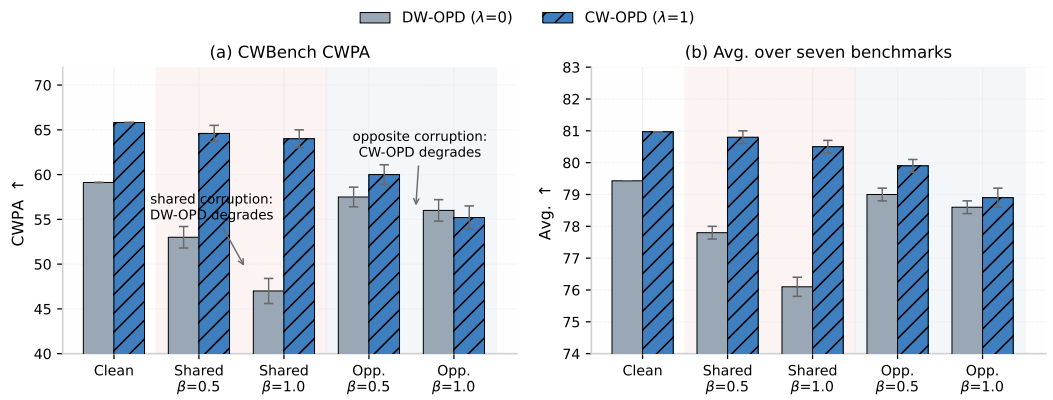

Figure 6: Robustness to structured teacher-logit perturbations. **Shared**: the same vector b is added in both worlds, under which the transition target is invariant (Proposition 1(i)); **opposite**: $+b/-b$ offsets the teacher contrast d_T by $2b$. Shared corruption degrades DW-OPD while CW-OPD is nearly unaffected; opposite corruption reverses the ordering, confirming that the two error structures act on the two objectives in opposite order. Error bars: std over 3 seeds.

that Proposition 1(i) is a statement about the transition target; shared perturbations can still reach CW-OPD through the endpoint loss, and the experiment is not claimed to verify the gradient-level statement (ii) or the stationary-point argument of Sec. 3.3. It confirms, at the level of trained models, that the two error structures act on the two objectives in opposite order.

D Result With Standard OPD

In the main text, CW-OPD is instantiated on top of on-policy self-distillation (OPSD), where the teacher is an EMA copy of the student. Here we verify that the proposed cross-world supervision is agnostic to the underlying distillation form by instantiating CW-OPD with a standard larger frozen teacher: Qwen3.5-9B teaches Qwen3.5-4B. Everything else, including the dual-world training pairs, the transition objective, and the training budget, is kept identical to the main experiments. Table 5

Table 5: Results with a standard larger teacher (Qwen3.5-9B → Qwen3.5-4B). CW-OPD (OPD) applies cross-world transition supervision on top of standard OPD. CWPA and CWFR are defined in Eqs. 9 and 10.

Method	V*Bench	HR-Bench 4K	RealWorldQA	MMVP	HalluB	SEED-B	OK-VQA	Avg.	CWPA ↑	CWFR ↓
Base	81.15	83.00	74.38	77.67	70.42	79.25	75.17	77.29	47.64	46.13
OPD	82.70	83.25	76.84	79.17	68.17	80.31	76.56	78.14	49.53	44.02
Vision-OPD	84.10	83.90	77.20	80.00	69.90	80.55	76.80	78.92	50.33	43.47
VAD	84.54	83.67	78.32	81.13	68.43	**80.79**	**78.93**	79.40	54.32	42.57
VA-OPD	85.69	84.32	76.29	78.67	72.34	79.96	78.33	79.37	54.27	40.96
FP-OPD	83.67	83.50	77.80	79.33	69.50	80.60	77.30	78.81	49.05	44.36
VGS	83.40	83.30	77.10	79.00	69.20	80.40	76.90	78.47	48.81	44.71
CW-OPD (OPD)	85.25	84.78	82.43	82.00	68.72	80.68	78.76	80.37	55.44	41.26

shows that the conclusions of the main text carry over to the standard-teacher setting: CW-OPD improves over other OPD baselines on both the standard benchmarks and, more substantially, on the cross-world diagnostic (CWPA and CWFR), despite the teacher and student now differing in capacity rather than in input conditions. This consistency indicates that the benefit of cross-world transition supervision does not rely on the privileged hint or on self-distillation, but stems from making the student's response to evidence changes an explicit optimization target—regardless of which teacher provides the per-world anchor.

E Training Setting and Details

We train Qwen3.5-4B and Qwen3.5-0.8B under the same on-policy distillation setup. The two configurations are identical except for the global batch size (48 for 4B and 64 for 0.8B), which we adjust to accommodate model scale and optimization stability. Both models use AdamW with a constant learning rate of 2×10^{-6}, weight decay 10^{-2}, and a single training epoch. The vision encoder is not frozen. On-policy rollouts use temperature 2.0 and top-p 1.0, with a maximum input prompt length of 8192 and a maximum response length of 1024.

For OPSD, the teacher is an exponential moving average (EMA) of the student rather than a fixed external model. The EMA teacher is initialized as a frozen copy of the student's initial weights and updated after every optimizer step as

$$\theta_{\text{teacher}} \leftarrow (1-\gamma)\,\theta_{\text{teacher}} + \gamma\,\theta_{\text{student}},$$

with $\gamma = 0.05$, corresponding to an EMA decay of 0.95. Only parameters are updated; buffers are not copied. This setting is used for both model sizes. For CW-OPD, we set the transition weight $\lambda = 1.0$. Tables 6 and 7 summarize the full configurations. We train with FSDP and vLLM under Python 3.12, PyTorch 2.10.0, Transformers 5.5.0, vLLM 0.18.0, and Ray 2.53.0. Rollout uses eight workers with tensor parallelism 1. The perception chat template prepends an empty ¡think¿¡/think¿ block; it does not request or super- vise a reasoning trace. For OPSD, the 4B run requires 2.34 hours(82 PFLOPs) and 0.8b run requires 1.59 hours(15 PFLOPs) on two NVIDIA RTX PRO 6000D GPUs (96 GB). For OPD, the 4B run requires 3.07 hours(108 PFLOPs) and 0.8b run requires 2.78 hours(36.2 PFLOPs) on two NVIDIA RTX PRO 6000D GPUs (96 GB).

Consistency across compared methods. All methods reported in this work share the same base training configuration described above, including the optimizer, learning rate, weight decay, schedule, epoch count, batch size, and on-policy rollout parameters. The standard OPD baseline (e.g.,VGS, FP-OPD, Vision-OPD) and all OPSD-style methods (e.g., RP-OPSD) use the same EMA teacher setting with update rate $\gamma = 0.05$ and per-step updates, so that the teacher signal is generated under an identical protocol. Only method-specific hyperparameters are varied, such as the transition weight λ in CW-OPD or the method-specific coefficients in VGS and related approaches. This ensures that performance differences reflect the distillation objective rather than differences in the training or teacher-update setup.

F Ablation of λ

We ablate the transition weight λ in the CW-OPD objective

$$\mathcal{L}_{\text{CW-OPD}} = \mathcal{L}_{\text{world}} + \lambda \mathcal{L}_{\text{trans}}$$

on the 4B model under both OPD and OPSD settings. Here $\lambda = 0$ reduces to bidirectional OPD without cross-world transition supervision. Each entry is the average over 3 seeds.

The local analysis of Sec. 3.3 makes only a qualitative prediction for this sweep. At a shortcut stationary point, a small λ decreases $\mathcal{T}$ to first order while $\mathcal{E}$ changes only at $O(\eta^2)$ (Eq. equation 8), so early in training—when the endpoint gradient has not yet vanished—a small transition weight adds an almost free corrective signal. As λ grows, $\mathcal{J}_\lambda$ becomes dominated by the transition loss and the endpoint fit may be sacrificed. The analysis does not guarantee an interior optimum or any specific shape of the performance curve; the existence and location of the peak are empirical questions, which Table 8 answers. The sweep confirms the qualitative picture: performance rises from $\lambda = 0$ to a peak at moderate λ (1.0 for OPD and 0.25–1.0 for OPSD) and declines as λ increases further, while remaining above the $\lambda = 0$ baseline throughout. We therefore use $\lambda = 1.0$ in all main experiments.

G Standard Benchmark Details

Beyond CW-Bench, we evaluate on eight standard benchmarks that we group into three categories according to what they primarily measure: visual perception and grounding, hallucination diagnosis, and external-knowledge reasoning. Table 9 summarizes the key properties of each benchmark.

Table 6: Key hyperparameters used for Qwen3.5-4B.

Hyperparameter	Value
Optimization & Training	
Optimizer	AdamW
Learning Rate	2e-6
Weight Decay	1e-2
LR Schedule	Constant
Epochs	1
Global Batch Size	48
Freeze Vision Encoder	False
On-Policy Rollout	
Rollout Temperature	2.0
Rollout Top-p	1.0
Max Input Prompt Length	8192
Max Response Length	1024
EMA Teacher (OPSD)	
Teacher Update Rate (γ)	0.05
EMA Decay	0.95
Update Interval	Every optimizer step
Teacher Initialization	Frozen copy of student
Buffer Update	No
CW-OPD Specific (Ours)	
λ	1.0

Table 7: Key hyperparameters used for Qwen3.5-0.8B.

Hyperparameter	Value
Optimization & Training	
Optimizer	AdamW
Learning Rate	2e-6
Weight Decay	1e-2
LR Schedule	Constant
Epochs	1
Global Batch Size	64
Freeze Vision Encoder	False
On-Policy Rollout	
Rollout Temperature	2.0
Rollout Top-p	1.0
Max Input Prompt Length	8192
Max Response Length	1024
EMA Teacher (OPSD)	
Teacher Update Rate (γ)	0.05
EMA Decay	0.95
Update Interval	Every optimizer step
Teacher Initialization	Frozen copy of student
Buffer Update	No
CW-OPD Specific (Ours)	
λ	1.0

G.1 Visual Perception and Grounding

V*Bench (Wu & Xie, 2024) evaluates whether MLLMs can process high-resolution, visually crowded images and locate small details. Unlike benchmarks centered on large, salient objects, V*Bench requires the model to find and focus on a specific visual key in a cluttered scene. It contains 191 multiple-choice questions across two tasks: attribute recognition (115 samples, 4 options) and spatial relationship reasoning (76 samples, 2 options). The benchmark was introduced alongside

Table 8: Ablation of the transition weight λ on the 4B model. $\lambda = 0$ corresponds to bidirectional OPD without transition supervision. Each entry is the average over 3 seeds.

λ	OPD + CW-OPD	OPSD + CW-OPD
base	77.29	77.29
0.0	77.81	77.73
0.25	79.88	80.46
0.5	80.07	80.32
1.0	80.37	80.97
2.0	79.64	79.39
4.0	78.31	77.62

Table 9: Overview of the eight standard benchmarks, organized by evaluation focus. "Task" denotes the primary question format; "Scale" reports the approximate number of evaluation samples or questions; "Metric" is the primary evaluation metric.

Category	**Benchmark**	**Task**	**Scale**	**Metric**
Visual Perception & Grounding	V*Bench (Wu & Xie, 2024)	MCQ	191 Qs	Accuracy
	HR-Bench 4K (Wang et al., 2024)	MCQ	4K images	Accuracy
	RealWorldQA	MCQ	765 Qs	Accuracy
	MMVP (Tong et al., 2024)	MCQ	300 Qs	Accuracy
Hallucination Diagnosis	POPE (Li et al., 2023b)	Binary QA	3 subsets	Acc / F1
	HallusionBench (Guan et al., 2024)	Paired QA	1,129 Qs	Pair Acc
External-Knowledge Reasoning	SEED-Bench (Li et al., 2023a)	MCQ	19K Qs	Accuracy
	OK-VQA (Marino et al., 2019)	Open QA	14K Qs	Soft Acc

the V* visual search algorithm, but the evaluation protocol itself measures raw MLLM performance without requiring the search mechanism.

HR-Bench 4K (Wang et al., 2024) is the first benchmark deliberately designed to assess MLLM perception on true 4K-resolution images. It contains images at 4K (and 8K in a separate split) paired with multiple-choice questions, covering fine-grained single-instance perception and cross-object perception. The benchmark reveals a substantial gap between model and human performance: state-of-the-art MLLMs achieve approximately 63% accuracy while humans reach 87%. We use the 4K split (HR-Bench 4K) in our main experiments.

RealWorldQA is a general visual question answering benchmark that tests fundamental visual perception capabilities in real-world image contexts. It consists of multiple-choice questions spanning diverse everyday scenes and object configurations, providing a broad measure of a model's basic visual understanding without additional domain-specific knowledge requirements.

MMVP (Tong et al., 2024) exposes systematic visual shortcomings in multimodal LLMs through "CLIP-blind pairs": image pairs that CLIP perceives as similar despite clear visual differences. The benchmark consists of 300 multiple-choice questions spanning nine basic visual patterns (e.g., orientation, presence, color, count, position, size). Even GPT-4V struggles on these questions, often providing incorrect answers and hallucinated explanations. MMVP thus isolates failures rooted in the visual representation itself rather than in language reasoning.

G.2 Hallucination Diagnosis

POPE (Li et al., 2023b) (Polling-based Object Probing Evaluation) assesses object hallucination through targeted binary queries: for each image, the model is asked whether a specific object is present. The benchmark aggregates data from three distinct sources (MSCOCO, A-OKVQA, and GQA) and evaluates under three sampling settings—random, popular, and adversarial—to probe how object frequency and co-occurrence bias affect hallucination rates. Performance is measured using accuracy and F1 score.

HallusionBench (Guan et al., 2024) is a comprehensive diagnostic suite for image-context reasoning that specifically targets *entangled* language hallucination and visual illusion. It comprises 346

images paired with 1,129 carefully crafted questions organized into control groups, enabling quantitative analysis of response tendencies, logical consistency, and failure modes. The paired-question structure allows the benchmark to measure whether a model's answer is grounded in the visual context or driven by language priors. GPT-4V achieves only 31.42% question-pair accuracy, with all other evaluated models below 16%.

G.3 External-Knowledge Reasoning

SEED-Bench (Li et al., 2023a) is a large-scale benchmark for generative comprehension in MLLMs. It contains 19K multiple-choice questions with human annotations, spanning 12 evaluation dimensions that cover both image and video comprehension. The questions are generated through a pipeline combining automatic filtering and manual verification, and the multiple-choice format with ground-truth options enables objective and efficient evaluation without human or GPT intervention during scoring. SEED-Bench measures a broad range of comprehension abilities, from single-image understanding to temporal reasoning in video.

OK-VQA (Marino et al., 2019) is a knowledge-based visual question answering benchmark specifically designed so that each question *requires external knowledge* beyond what is visible in the image. It consists of approximately 14,000 images and 14,000 open-ended questions sampled from MS-COCO, spanning 11 high-level knowledge categories including science, sports, geography, culture, and history. Each question is paired with multiple human-provided answers to enable consensus-based evaluation. OK-VQA probes a model's ability to (i) extract relevant visual cues, (ii) retrieve relevant external knowledge, and (iii) integrate perception with knowledge to produce a coherent answer. Early VQA architectures performed near random chance on OK-VQA, and even modern models struggle with rare or niche knowledge categories.

G.4 Evaluation settings.

All model responses are scored using Qwen3.5-9B as the LLM judge under greedy decoding: temperature 0, top-p 1, and a maximum generation length of 64 tokens. Thinking mode is disabled (`enable_thinking=false`), and the judge runs with tensor parallelism 2. Rule-based matching is applied first where applicable; the LLM judge is used only for unparsed or open-ended responses.

LLM Judge Prompt Template

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