VisionRL-SD: Stabilizing Visual On-Policy Self-Distillation with Verifiable Spatial Rewards

Anonymous ACL submission

Abstract

On-policy self-distillation (OPSD), in which a single model serves as both teacher and student under different conditioning, has emerged as a key post-training paradigm. Existing methods inject the teacher's advantage through additional text in its prompt, while teacher and student share the same visual input. A natural counterpart for vision-language models places this advantage in the visual channel by overlaying annotations on the teacher's image, yet the training dynamics of OPSD in this visual setting have not been systematically studied. We provide such a study on a controlled spot-the-difference task, where a vision-language model must localize the difference between two similar images with a bounding box. We observe that pure visual OPSD is unstable, and that removing the outcome-level reward collapses validation IoU to near zero. We address this by combining group relative policy optimization using an IoU reward with token-level distillation restricted to trajectories of intermediate spatial quality. Through extensive ablations we identify three design choices that matter most: the reward-distillation weight follows an inverted-U; band-pass selective distillation outperforms uniform distillation; and a repetition penalty is the single most impactful auxiliary reward component. On Qwen3.5-2B, our method improves IoU@0.5 from 56.6 to 63.7.

1 Introduction

Post-training of large language models has increasingly relied on reinforcement learning with verifiable rewards (RLVR) to improve reasoning, alignment, and structured prediction (Shao et al., 2024; DeepSeek-AI et al., 2025; Kimi Team et al., 2025). In parallel, on-policy distillation has emerged as a complementary paradigm that provides dense token-level supervision on the model's own rollouts, mitigating the train-inference distribution mismatch of off-policy distillation (Agarwal et al., 2024). A particularly attractive instantiation is on-policy self-distillation (OPSD), where a single model serves as both teacher and student under different conditioning, eliminating the need for a separately maintained teacher (Zhao et al., 2026; Hübotter et al., 2026; Penaloza et al., 2026; Shenfeld et al., 2026). Recent work has further shown that reinforcement learning and token-level distillation are not competing paradigms but provide complementary signals, with verifiable rewards anchoring outcome quality while distillation delivers per-token credit assignment (Xu et al., 2025; Yang et al., 2026; Ding, 2026). Existing OPSD methods have, however, developed almost entirely in the text domain. The teacher's advantage over the student is typically injected through additional text in its conditioning context, such as reference solutions (Zhao et al., 2026), environment feedback (Hübotter et al., 2026), or expert demonstrations (Shenfeld et al., 2026). The student and teacher share the same visual or environmental input, and the asymmetry is purely textual. A natural counterpart for vision-language models places this advantage in the visual channel: the teacher's image is overlaid with annotations indicating the target region, while the student receives only the raw image. This formulation is particularly well-aligned with spatial localization tasks, where the answer is difficult to express compactly in text but trivial to mark in the image. Despite this natural fit, the training behavior of OPSD in the visual setting has not, to our knowledge, been systematically studied. We provide such a study using a controlled spot-the-difference task, in which a vision-language model must localize the difference between two similar images with a bounding box. The teacher receives the image pair with the difference region annotated; the student receives the raw pair. Recent text-only OPSD studies have reported a characteristic rise-then-fall training pattern, where performance initially improves but deteriorates as optimization

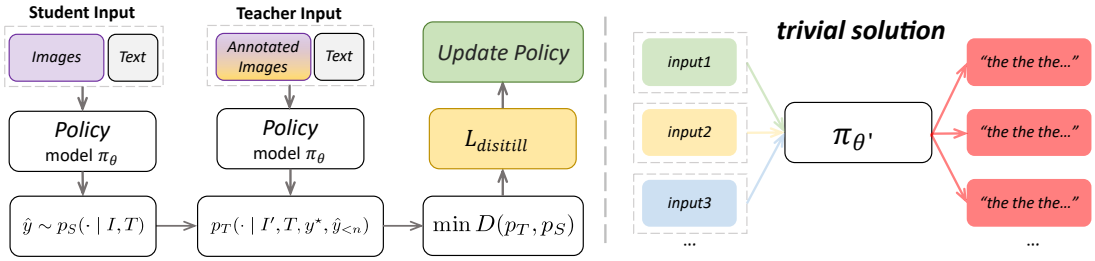


Figure 1: **Left:** Pure visual on-policy self-distillation (OPSD): a shared-weight policy serves as both student (receiving raw images) and teacher (receiving annotated images), with training minimizing the token-level divergence $D(p_T, p_S)$. **Right:** This objective admits a trivial solution—an input-invariant policy that produces the same degenerate output regardless of input, achieving zero divergence by ignoring all visual information. While training instability in text-only OPSD manifests as gradual degradation (Luo et al., 2026; Fu et al., 2026; Li et al., 2026), we observe a far more severe collapse in the visual setting: validation IoU drops to near zero within the first few steps (Section 3).

proceeds, often accompanied by output length inflation and n-gram repetition (Luo et al., 2026; Fu et al., 2026; Li et al., 2026). We find that this issue is substantially more severe in the visual setting: under the pure visual OPSD objective, i.e., without an outcome-level reward, validation IoU collapses to near zero within the first few training steps. This indicates that token-level distribution matching alone is insufficient when the teacher's advantage resides in the visual input while the distillation loss acts on language tokens. We bring the hybrid distillation-RL recipe to this visual setting and ask how it should be configured. We anchor visual OPSD with group relative policy optimization (GRPO) (Shao et al., 2024) using an IoU-based reward, and restrict token-level distillation to trajectories of intermediate spatial quality, motivated by the observation that the shared-parameter teacher is unreliable on severely off-track rollouts and redundant on already-correct ones. The selectivity principle has appeared in text-only work in different forms (Ding, 2026; Ke et al., 2026); we provide its instantiation for visual localization through IoU-based quality bands. Our contributions are summarized as follows:

- We provide a systematic study of OPSD in the visual setting, showing that pure visual OPSD is unstable and that outcome-level rewards are necessary.
- We adapt hybrid distillation-RL training to visual self-distillation and identify three key design choices through extensive ablations.
- We release a synthetic spot-the-difference dataset as a diagnostic testbed, improving IoU@0.5 from 56.6 to 63.7 on Qwen3.5-2B and from 61.1 to 64.6 on Qwen3.5-4B.

2 Related Work

On-Policy Self-Distillation in the Text Domain. On-policy distillation trains a student on its own rollouts under dense token-level supervision from a teacher, mitigating the train-inference distribution mismatch of off-policy distillation (Agarwal et al., 2024). A growing line of work replaces the external teacher with the model itself under privileged conditioning, giving rise to on-policy self-distillation (OPSD), where the teacher's advantage is injected through additional text in its prompt: reference solutions (Zhao et al., 2026), environment feedback such as runtime errors (Hübotter et al., 2026), privileged reasoning traces in agentic environments (Penaloza et al., 2026), or expert demonstrations (Shenfeld et al., 2026). Several recent studies report failure modes shared across these text-domain instantiations, including abrupt length inflation, repetition saturation, and unreliable teacher guidance on student-generated prefixes (Luo et al., 2026; Fu et al., 2026; Li et al., 2026). All of these works operate in the text channel and leave open how OPSD behaves when the teacher's advantage is placed in the visual input.

Hybrid Distillation-RL and Selective Supervision. A parallel line of work observes that reinforcement learning with verifiable rewards and token-level distillation provide complementary signals: rewards anchor outcome quality while distillation delivers per-token credit assignment. KDRL unifies GRPO with reverse-KL distillation under a single objective (Xu et al., 2025), RLSD uses

environment rewards to determine update direction and a privileged teacher to set per-token magnitude (Yang et al., 2026), and HDPO targets distillation at cliff prompts where all rollouts fail and the RL gradient vanishes (Ding, 2026). A related thread shows that uniform token-level supervision is suboptimal and that gating by reliability or uncertainty improves stability (Ke et al., 2026). These works share our high-level recipe of combining outcome-level rewards with selective token-level distillation, but operate exclusively on textual reasoning; the visual instantiation we study presents a different reliability signal (IoU) and a different failure mode (collapse without outcome anchoring).

Vision-Language Models for Spatial Localization and Change Understanding. Vision-language models built on visual instruction tuning have advanced multimodal understanding (Liu et al., 2023; Chen et al., 2023; OpenAI, 2023; Yang et al., 2023), yet precise spatial localization remains challenging (Kazemzadeh et al., 2014; Yu et al., 2016; Liu et al., 2024). Reinforcement learning with verifiable rewards has been applied to VLM grounding by treating IoU as a reward signal (Shao et al., 2024; DeepSeek-AI et al., 2025), but most such efforts use RL alone without token-level teacher supervision. Spot-the-difference localization additionally requires comparing two similar images and identifying a localized change (Jhamtani and Berg-Kirkpatrick, 2018; Park et al., 2019; Alcantarilla et al., 2018), a setting largely addressed with task-specific architectures rather than general-purpose VLMs.

3 Task and Dataset

We study visual OPSD on a spot-the-difference localization task. Given a pair of similar images (I_1, I_2) that differ in exactly one region, the model must output a bounding box $[x_1, y_1, x_2, y_2]$ localizing the changed area. This task is well-suited to our study for two reasons. First, it naturally provides visual privileged information: the teacher's images can be annotated with bounding boxes highlighting the difference region, placing the teacher's advantage entirely in the visual channel. Second, it admits a continuous, verifiable reward (IoU) that enables both reinforcement learning and quality-based gating of the distillation loss.

Synthetic Dataset. To enable controlled study, we construct a synthetic spot-the-difference dataset.

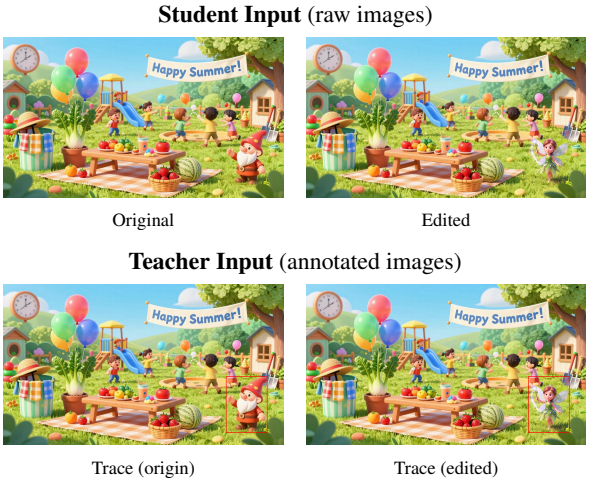
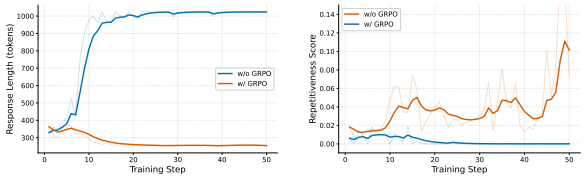


Figure 2: A spot-the-difference example. The student receives raw image pairs; the teacher receives the same images annotated with bounding boxes highlighting the difference region (privileged information, unavailable at inference).

Synthetic generation provides exact ground-truth bounding boxes and pixel-aligned annotated images for the teacher, while affording full control over edit types and difficulty. Since the task requires only detecting localized differences, photo-realism is unnecessary. Full pipeline details are in Appendix D.

Pipeline. The dataset is generated in five stages. (1) A language model generates diverse scene descriptions guided by a WordNet-derived taxonomy of 25 object categories (Fellbaum, 1998). (2) A text-to-image diffusion model renders each scene at 1344×768 resolution in one of eight artistic styles, including cartoon, anime, watercolor, and comic-book. (3) An open-vocabulary detector (Liu et al., 2024) identifies candidate objects and their bounding boxes. (4) The language model plans an edit by selecting a target object and specifying a modification. (5) An image editing model applies the edit, with Poisson blending (Pérez et al., 2003) for seamless integration. Each edit produces an original–edited image pair together with an annotated visualization of the difference region used as the teacher's privileged input.

Edit Types and Statistics. The pipeline supports three edit types: *replacement*, *color change*, and *removal*. Each pair contains exactly one edit. The final dataset contains 1,142 image pairs spanning diverse themes across eight artistic styles; we use 400 for training and 742 for validation. This scale suffices for studying visual OPSD training dynam-



(a) Response length (b) Repetitiveness

Figure 3: Training dynamics: pure visual OPSD (w/o GRPO) vs. the hybrid objective (VisionRL-SD). Without outcome-level rewards, the model collapses within the first few training steps.

ics, as both GRPO and the hybrid objective converge within a few hundred steps.

Instability of Pure Visual OPSD. Before introducing our method, we examine the baseline behavior of pure visual OPSD on this task. We train Qwen3.5-2B under the OPSD objective alone—minimizing the token-level divergence between the teacher (receiving annotated images) and the student (receiving raw images)—without any outcome-level reward. As shown in Figure 3, the model collapses within the first few training steps: response length saturates at the maximum token limit and n-gram repetition escalates sharply, while gradient norms remain bounded—indicating an optimization pathology rather than numerical instability. Validation IoU drops to near zero, confirming that the model has learned to suppress its dependence on the visual input rather than to localize the difference region. This collapse is substantially more severe than the gradual rise-then-fall patterns reported in text-only OPSD (Luo et al., 2026; Fu et al., 2026; Li et al., 2026), confirming that outcome-level anchoring is essential when the teacher's advantage resides in the visual channel.

4 Method

Motivated by the instability observed in Section 3, we now formalize the privileged on-policy self-distillation (OPSD) setting in the visual channel and present a hybrid training framework that anchors token-level distillation with outcome-level reinforcement learning. Section 4.1 states the problem, Section 4.2 introduces the hybrid objective, Section 4.3 describes the band-pass selective distillation, and Section 4.4 specifies the reward design.

4.1 Problem Setup

Let π_θ denote a vision-language model parameterized by θ. Given a prompt p and an image pair (I_1, I_2), the model generates a response y autoregressively:

$$\pi_\theta(y \mid p, I_1, I_2) = \prod_{t=1}^{T} \pi_\theta(y_t \mid p, I_1, I_2, y_{<t}). \quad (1)$$

On the spot-the-difference task considered in this work, the response contains a bounding box $b = [x_1, y_1, x_2, y_2]$ localizing the difference region between I_1 and I_2.

Visual Privileged Self-Distillation. In privileged OPSD, a single set of parameters θ serves as both teacher and student under different conditioning. Existing methods inject the teacher's advantage through additional text in its prompt (Zhao et al., 2026; Hübotter et al., 2026; Penaloza et al., 2026; Shenfeld et al., 2026); here we place the advantage in the visual channel. The teacher receives an annotated image pair $(\tilde{I}_1, \tilde{I}_2)$ in which the ground-truth difference region is overlaid with a red bounding box, while the student receives the raw pair (I_1, I_2):

$$x_{\text{tea}} = (p, \tilde{I}_1, \tilde{I}_2), \qquad x_{\text{stu}} = (p, I_1, I_2). \quad (2)$$

The teacher forward pass is computed under `no_grad()`, providing a frozen snapshot of the current policy at each training step. The goal is for the student to recover the annotated region from the raw view alone, internalizing the spatial information that the teacher reads directly off its annotated input.

Degenerate Solution of Pure Visual OPSD. This visual formulation makes pure self-distillation ill-posed without an outcome-level anchor. The teacher's advantage is provided through the visual input, whereas the distillation loss is applied only to token distributions. Ignoring the rollout prefix for notation, the pure OPSD objective minimizes

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_y \left[D\big(\pi_\theta(\cdot \mid x_{\text{tea}}) \,\|\, \pi_\theta(\cdot \mid x_{\text{stu}})\big) \right]. \quad (3)$$

Since the teacher and student share parameters, any input-invariant policy $\pi_{\theta^\star}(\cdot \mid x) = q(\cdot)$ makes the two conditional distributions identical and attains zero divergence. The loss therefore cannot distinguish a student that has learned to localize the raw image from one that has simply suppressed its dependence on the visual input. This issue is especially severe in the visual privileged setting because the privileged signal is provided in the visual channel, while the supervision acts only on language-token distributions. Matching the teacher's token

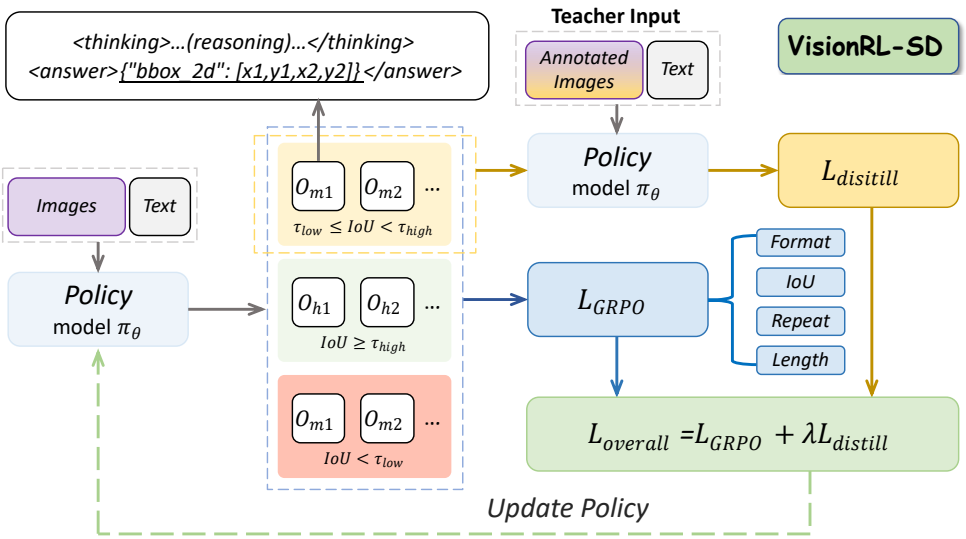


Figure 4: Overview of VisionRL-SD. The student generates rollouts from raw images, which are scored by IoU and sorted into three quality bands. Only rollouts in the intermediate band ($\tau_{\text{low}} \leq \text{IoU} < \tau_{\text{high}}$) receive token-level distillation from the teacher (which receives annotated images as privileged input). All rollouts receive GRPO updates with a composite reward (IoU, format, length, repetition). The combined loss $\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{GRPO}} + \lambda \mathcal{L}_{\text{distill}}$ updates the shared policy.

distribution therefore does not directly require recovering the spatial grounding path from the raw image.

4.2 Hybrid Distillation-RL Objective

Motivated by the degeneracy above and the empirical collapse observed in Section 3, we introduce a reinforcement learning objective with verifiable spatial rewards as an outcome-level anchor. Following the hybrid distillation-RL recipe established in text-only work (Xu et al., 2025; Yang et al., 2026; Ding, 2026), we combine token-level distillation with GRPO.

The training objective is a weighted sum of an RL term and a token-level distillation term:

$$\mathcal{L}(\theta) = \lambda_{\text{grpo}} \mathcal{L}_{\text{GRPO}}(\theta) + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}}(\theta). \quad (4)$$

Outcome-Level Reward via GRPO. We adopt group relative policy optimization (GRPO) (Shao et al., 2024; DeepSeek-AI et al., 2025) as the RL component. For each prompt, n rollouts $\{y_i\}_{i=1}^{n}$ are sampled from $\pi_{\theta_{\text{old}}}$, and the group-normalized advantage is computed as $\hat{A}_i = (r_i - \bar{r})/\sigma_r$, where $\bar{r}$ and σ_r are the mean and standard deviation of rewards within the group. The same scalar advantage is shared across all tokens of a response, yielding the clipped surrogate

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\Big[\sum_{t=1}^{T} \min\big(\rho_t \hat{A}, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\hat{A}\big)\Big], \quad (5)$$

with per-token importance ratio $\rho_t = \pi_\theta(y_t \mid \cdot)/\pi_{\theta_{\text{old}}}(y_t \mid \cdot)$. The GRPO loss is applied to all rollouts and provides a global signal that directly penalizes spatially incorrect or degenerate outputs.

Token-Level Distillation. The distillation term minimizes a token-level divergence between teacher and student conditional distributions over the student's own rollouts:

$$\mathcal{L}_{\text{distill}}^{\text{full}} = \mathbb{E}_{y \sim \pi_\theta}\Big[\sum_{t=1}^{T} D\big(\pi_\theta(\cdot \mid x_{\text{tea}}, y_{<t}) \,\|\, \pi_\theta(\cdot \mid x_{\text{stu}}, y_{<t})\big)\Big]. \quad (6)$$

We use Jensen–Shannon divergence with a top-k approximation: the divergence is computed exactly over the k most probable tokens under the teacher distribution and aggregated into a single tail bucket otherwise, reducing the per-step cost from $O(|V|)$ to $O(k)$ while preserving the information-dense head of the distribution. We ablate the choice of divergence and the top-k approximation in Section 5.

4.3 Band-Pass Selective Distillation

Applying token-level distillation uniformly to all rollouts is suboptimal. We identify two regimes

in which uniform distillation is ineffective and introduce a selective gating mechanism that restricts distillation to an intermediate quality band. This follows the broader principle, also articulated in text-only OPSD work, that uniform supervision should be replaced by gating based on a reliability or uncertainty signal (Ding, 2026; Ke et al., 2026); we instantiate this principle for visual localization through IoU-based quality bands.

Two Regimes of Diminished Returns. At high spatial quality, the rollout already lies close to the optimum and the RL reward provides a strong, well-shaped gradient; the additional token-level signal from the teacher is largely redundant. At low spatial quality, the rollout has drifted far from any useful path, and because teacher and student share weights, the teacher's conditional distribution over such an off-track prefix is itself unreliable and can mislead the student (Ross et al., 2011; Agarwal et al., 2024).

Quality-Band Gating. We restrict the distillation loss to rollouts whose IoU falls inside an intermediate band $[\tau_{\text{low}}, \tau_{\text{high}})$:

$$m(y) = \mathbb{1}\left[\tau_{\text{low}} \leq \text{IoU}(y) < \tau_{\text{high}}\right], \qquad (7)$$

yielding the band-pass selective distillation loss

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{y\sim\pi_\theta}\Bigg[m(y) \cdot \sum_{t=1}^{T} D_{\text{JSD}}\big(\pi_\theta(\cdot \mid x_{\text{tea}}, y_{<t}) \parallel \pi_\theta(\cdot \mid x_{\text{stu}}, y_{<t})\big)\Bigg]. \qquad (8)$$

The gating uses raw IoU rather than the composite reward defined below, so that gating quality is isolated from format and length effects. Importantly, the GRPO loss in Eq. 4 continues to be applied to *all* rollouts regardless of m: only the distillation term is gated.

4.4 Composite Reward

The scalar reward used by GRPO combines spatial accuracy with auxiliary terms that penalize degenerate outputs:

$$r(y, b^*) = r_{\text{iou}} + r_{\text{format}} - r_{\text{len}} - r_{\text{rep}}, \qquad (9)$$

where $r_{\text{iou}} = \text{IoU}(b_{\text{pred}}, b^*)$ is the intersection-over-union between the predicted and ground-truth bounding boxes, r_{format} rewards structured chain-of-thought outputs containing well-formed `<thinking>` and `<answer>` tags with a valid bounding box, $r_{\text{len}} \geq 0$ linearly penalizes outputs exceeding a character-length threshold, and $r_{\text{rep}} \geq 0$ penalizes character-level n-gram repetition above a threshold τ_{rep}. The IoU term drives spatial accuracy, while the penalty terms suppress length inflation and repetition patterns that token-level distillation alone would otherwise reinforce. We analyze the relative importance of each component in Section 5 and find that the repetition penalty in particular is the single most impactful auxiliary reward component.

Algorithm 1 Training Loop for Reward-Anchored Visual OPSD

Require: Policy π_θ, dataset $\mathcal{D}$, band thresholds $\tau_{\text{low}}, \tau_{\text{high}}$, weights $\lambda_{\text{grpo}}, \lambda_{\text{distill}}$

for each training step **do**
  Sample batch $\{(p_i, I_1^i, I_2^i, b^{*i})\}$ from $\mathcal{D}$
  Generate rollouts $y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid p_i, I_1^i, I_2^i)$
  Compute rewards r_i and IoU scores s_i for each y_i
  Compute teacher logits under $(\tilde{I}_1^i, \tilde{I}_2^i)$ with `no_grad`
  Band-pass filter: $\mathcal{B} \leftarrow \{i : \tau_{\text{low}} \leq s_i < \tau_{\text{high}}\}$
  $\mathcal{L}_{\text{GRPO}} \leftarrow$ clipped policy gradient over all samples
  $\mathcal{L}_{\text{distill}} \leftarrow$ top-k JSD over samples in $\mathcal{B}$
  Update θ by $\nabla_\theta(\lambda_{\text{grpo}}\mathcal{L}_{\text{GRPO}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}})$
end for

4.5 Training Procedure

Algorithm 1 summarizes the full training loop. Because the policy is updated between rollout generation and the distillation loss computation, the teacher logits used at the loss step are drawn from a slightly stale distribution; we apply an importance sampling correction to account for this shift (details in Appendix A).

5 Experiments

5.1 Experimental Setup

Base Model. We use Qwen3.5-2B as our base vision-language model. For reference comparisons in Table 1, we evaluate Qwen3.5-9B, Qwen3.6-Plus, and GPT-5.5 under identical prompts and parsing pipelines; these serve as upper-bound references rather than controlled baselines.

Training Configuration. We train with GRPO using $n = 8$ rollouts per prompt with group-normalized advantages. Key hyperparameters include learning rate 1×10^{-6}, policy clip ratio $\epsilon = 0.2$, $\lambda_{\text{grpo}} = 1.0$, $\lambda_{\text{distill}} = 1.0$, band-pass thresholds $\tau_{\text{low}} = 0.3$ and $\tau_{\text{high}} = 0.5$, and top-k JSD distillation with $k = 100$. We use FSDP2 with gradient checkpointing on 2 GPUs and vLLM for

Method	IoU@0.5 ↑	Mean IoU ↑
Qwen3.6-Plus	84.1	0.7693
GPT-5.5	73.6	0.6254
Qwen3.5-9B	69.3	0.6247
CapRL-InternVL3.5-8B	3.6	0.1404
PRCO-8B	1.8	0.1216
Vero-Qwen3I-8B	4.3	0.1338
Qwen3.5-2B	56.6	0.5121
Qwen3.5-2B-SFT	62.1	0.5829
Qwen3.5-2B-EMA	60.7	0.5710
VisionRL-SD-2B	63.7 (+7.1)	0.5928 (+0.081)
Qwen3.5-4B	61.1	0.5512
VisionRL-SD-4B	64.6 (+3.5)	0.6173 (+0.064)

Table 1: Main results on the validation set.

rollout generation. Full hyperparameter details are in Appendix B.

Evaluation Metrics. All evaluations use greedy decoding (temperature $= 0$) for near-deterministic, reproducible results. We evaluate models using:

- **IoU@0.5:** Percentage of predictions with IoU ≥ 0.5
- **Mean IoU:** Average IoU across all predictions (invalid parses receive IoU $= 0$)

5.2 Main Results

Table 1 presents main results on the validation set, comparing VisionRL-SD against zero-shot baselines and proprietary reference models.

VisionRL-SD improves small models over zero-shot baselines. On Qwen3.5-2B, VisionRL-SD achieves +7.1 IoU@0.5 and +0.081 Mean IoU over the zero-shot baseline, a 12.5% relative improvement in IoU@0.5. On Qwen3.5-4B, the gain is +3.5 IoU@0.5 (+5.7% relative). Relative to the GRPO-only baseline (62.3%), adding band-pass distillation yields a further +1.4 IoU@0.5, which is consistent across ablation settings (Table 2). The larger improvement on the 2B model suggests that self-distillation provides greater benefit when the model's intrinsic capacity is more limited, as the privileged teacher signal compensates for weaker visual reasoning.

Discussion of baselines. We compare against zero-shot VLM baselines to measure the absolute gain from our training framework. We additionally include an EMA-teacher distillation baseline, where the teacher is a separate exponential moving average copy of the student parameters rather than the student itself. This baseline does not suffer from the shared-weight instability described in Section 3, and serves as a reference for what standard distillation achieves without the shared-weight constraint.

Configuration	IoU@0.5 ↑	Mean IoU ↑
VisionRL-SD-2B	63.7	0.5928
Training loss		
w/o GRPO loss	0.0	0.0
w/o self-distillation	62.3 (−1.4)	0.5840 (−0.009)
Reward shaping		
w/o length penalty	62.8 (−0.9)	0.5801 (−0.013)
repeat θ=0.2 (vs 0.4)	63.2 (−0.5)	0.5835 (−0.009)
w/o repetition penalty	60.6 (−3.1)	0.5773 (−0.016)

Table 2: Ablation study.

τ_{low}	τ_{high}	IoU@0.5 ↑	Mean IoU ↑
GRPO only		62.3	0.5841
0.0	1.0	61.9 (−0.4)	0.5825 (−0.002)
0.1	0.9	61.7 (−0.6)	0.5753 (−0.009)
0.3	0.8	63.3 (+1.0)	0.5887 (+0.005)
0.5	0.8	61.9 (−0.4)	0.5816 (−0.003)
0.3	0.5	**63.7** (+1.4)	**0.5928** (+0.009)

Table 3: Band-pass threshold sensitivity. GRPO only (no distillation) is shown as a reference baseline.

5.3 Ablation Study

Table 2 isolates the contribution of each component.

Removing GRPO causes complete collapse (0.0), confirming that outcome-level reward is essential. Removing self-distillation causes a moderate drop (−1.4), showing the value of dense token-level guidance. Among reward components, the repetition penalty is most critical (−3.1), serving as a secondary defense against degenerate outputs.

5.4 Sensitivity Analysis

5.4.1 Band-Pass Threshold Selection

Table 3 varies the band-pass thresholds $(\tau_{\text{low}}, \tau_{\text{high}})$ that control which samples receive distillation.

The best band $[0.3, 0.5]$ achieves 63.7%, outperforming both GRPO only (62.3%, +1.4) and unfiltered distillation $[0, 1.0]$ (61.9%, +1.8). Notably, unfiltered distillation ($[0, 1.0]$) performs *worse* than GRPO alone, indicating that indiscriminate teacher supervision actively harms training—the noise from low- and high-quality extremes outweighs the benefit. Band-pass filtering recovers the value of distillation by restricting it to the intermediate quality range where teacher guidance is informative. Setting $\tau_{\text{low}} = 0.3$ excludes severely erroneous tra-

λ_{distill}	IoU@0.5 ↑	Mean IoU ↑
0.0 (RL only)	62.3	0.5841
0.5	63.2	0.5898
1.0	63.7	0.5928
1.5	62.9	0.5928
2.0	61.9	0.5859
2.5	59.6	0.5672

Table 4: Sensitivity to distillation weight.

jectories where the teacher's signal is unreliable, while $\tau_{\text{high}} = 0.5$ excludes high-quality samples where GRPO alone suffices. Raising the lower threshold to 0.5 narrows the band too aggressively, reducing the number of eligible samples and degrading performance back to 61.9%. The active token fraction over training for each threshold setting is shown in Appendix Figure 6.

5.4.2 Distillation Weight

Table 4 examines sensitivity to λ_{distill}.

Performance follows an inverted-U pattern, peaking at $\lambda_{\text{distill}} = 1.0$ (63.7%). At $\lambda_{\text{distill}} = 0$ (GRPO only), the model lacks dense teacher guidance, while excessive distillation ($\lambda_{\text{distill}} \geq 2.0$) overwhelms the RL signal and degrades sharply—at $\lambda_{\text{distill}} = 2.5$, performance drops to 59.6%, even below the GRPO-only baseline. Beyond final performance, Figure 5 reveals a striking difference in training dynamics: GRPO alone suffers a reward dip in early training (steps 5–20) before recovering, whereas moderate distillation (λ_{distill}=1.0) eliminates this dip entirely, with reward increasing monotonically from the first step. This suggests that the dense teacher signal provides effective early-stage guidance that steers the policy past the initial exploration phase where sparse RL rewards alone are insufficient.

We also varied the divergence type and distillation mode (Table 6 in Appendix); JSD with top-k performs best, though the spread is small ($\leq$1.0 point), indicating robustness to these choices. As shown in Section 3 and Figure 3, pure visual OPSD without GRPO collapses within the first few training steps, while VisionRL-SD with GRPO anchoring achieves stable convergence, reaching 63.7% IoU@0.5 compared to the GRPO-only plateau of 62.3%.

6 Conclusion

We presented a systematic study of on-policy self-distillation in the visual setting, using a spot-the-difference localization task as a controlled testbed.

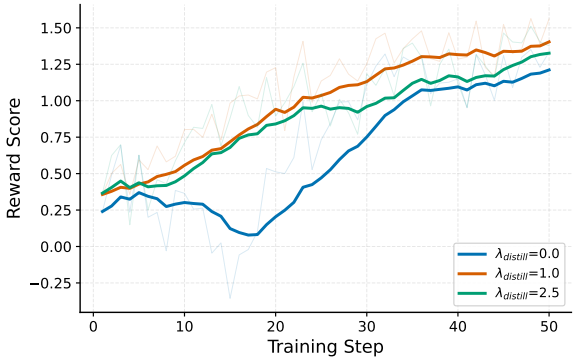


Figure 5: Reward over training for different λ_{distill} values. Moderate distillation (λ_{distill}=1.0) accelerates reward growth in early training compared to GRPO alone (λ_{distill}=0), while excessive distillation (λ_{distill}=2.5) may overwhelm the RL signal, leading to slower convergence.

We showed that pure visual OPSD is fundamentally unstable: without an outcome-level reward, the shared-weight divergence objective drives the model toward input-invariant solutions and validation IoU collapses to near zero within the first few training steps.

To address this, we adapted the hybrid distillation-RL recipe to visual self-distillation, anchoring token-level distillation with GRPO using an IoU-based reward and restricting distillation to trajectories of intermediate spatial quality. Through extensive ablations we identified three design choices that matter most: the reward-distillation weight follows an inverted-U, band-pass selective distillation outperforms uniform distillation, and a repetition penalty is the single most impactful auxiliary reward component. On Qwen3.5-2B, our method improves IoU@0.5 from 56.6 to 63.7; on Qwen3.5-4B, from 61.1 to 64.6.

Our observations are, in principle, applicable to any privileged self-distillation setting where teacher and student share weights but receive asymmetric inputs, provided a verifiable reward and a continuous quality metric are available. Candidate domains include visual grounding, agent tasks with ground-truth action sequences, and reasoning with solution hints; future work could also explore adaptive threshold scheduling and the equilibrium between distillation and RL gradients.

Limitations

Our evaluation has several limitations that should be considered when interpreting the results. First, we validate on a single task (spot-the-difference)

with synthetic images; while the instability of visual OPSD likely extends to other shared-weight asymmetric-input settings, empirical verification on other privileged OPSD domains remains future work. Constructing privileged information for other multimodal tasks is non-trivial: unlike spot-the-difference where annotated images naturally serve as privileged input, many vision-language tasks lack an obvious form of privileged information that can be overlaid on the input without changing the task semantics. Second, the band-pass thresholds are fixed throughout training; adaptive scheduling could improve efficiency as the model's quality distribution shifts. Third, our framework requires privileged annotations during training, which are trivially obtained in synthetic pipelines but may require additional effort for real-world data. Finally, our analysis of the degenerate optimum characterizes the loss landscape and empirical behavior on this task; a formal convergence proof under specific architectural assumptions remains open.

Ethics Statement

This work focuses on improving vision-language models for visual reasoning tasks. The synthetic dataset was generated using publicly available text-to-image models and does not contain personal information or sensitive content.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Pablo F. Alcantarilla, Simon Stent, Germán Ros, Roberto Arroyo, and Riccardo Gherardi. 2018. Street-view change detection with deconvolutional networks. *Auton. Robots*, 42(7):1301–1322.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 2 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645:633–638. ArXiv preprint arXiv:2501.12948.

Ken Ding. 2026. Hdpo: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. 2026. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*.

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

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4024–4034. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. 2026. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning. *arXiv preprint arXiv:2605.13255*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, Lecture Notes in Computer Science, pages 38–55. Springer.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. 2026. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*.

OpenAI. 2023. Gpt-4v(ision) system card. *OpenAI Technical Report*.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4623–4632. IEEE.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. 2026. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*.

Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. In *ACM SIGGRAPH 2003 Papers, San Diego, CA, USA, July 27-31, 2003*, pages 313–318. ACM.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, JMLR Proceedings, pages 627–635. JMLR.org.

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

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. 2025. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. 2026. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *CoRR*, abs/2309.17421.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, Lecture Notes in Computer Science, pages 69–85. Springer.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*.

A Importance Sampling Correction

Since the policy π_θ is updated between rollout generation and the distillation loss computation, a distribution shift arises between the rollout policy π_{old} and the current policy π_θ. We correct for this shift by reweighting the per-token distillation loss.

The importance weight for token t is:

$$w_t = \frac{\pi_\theta(y_t \mid \cdot)}{\pi_{\text{old}}(y_t \mid \cdot)} = e^{\log \pi_\theta(y_t|\cdot) - \log \pi_{\text{old}}(y_t|\cdot)} \quad (10)$$

We clip importance weights to prevent instability: $\tilde{w}_t = \min(w_t, c)$, where c is a clipping threshold. The IS-corrected distillation loss becomes:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{y\sim\pi_{\text{old}}}\Big[m_{\text{sel}}(y) \cdot \sum_{t=1}^{T} \tilde{w}_t \cdot D_{\text{JSD}}^{(k)}(\pi_{\text{tea}}(\cdot|\cdot),\, \pi_\theta(\cdot|\cdot))\Big] \quad (11)$$

We additionally apply a token-level IS correction at the rollout stage for the GRPO loss.

A.1 Top-k JSD Approximation

The top-k JSD restricts the divergence computation to the k most probable tokens under the teacher distribution. Let $\mathcal{T}_k$ denote the set of k tokens with highest teacher probability. The tail probability mass is aggregated into a single bucket:

$$P_{\text{tail}} = 1 - \sum_{v \in \mathcal{T}_k} P_{\text{tea}}(v) \quad (12)$$

$$Q_{\text{tail}} = 1 - \sum_{v \in \mathcal{T}_k} P_{\text{stu}}(v) \quad (13)$$

The JSD is then computed over $k + 1$ categories (the k top tokens plus the tail bucket), reducing vocabulary-sized computation to $O(k)$ per token position.

B Hyperparameter Details

Table 5 lists all hyperparameters used in training.

Hyperparameter	**Value**
Optimization	
Learning rate	1×10^{-6}
Warmup steps	10
Scheduler	Constant
Gradient clipping	1.0
Train batch size	1
Rollout batch size	4
Rollouts per prompt	8
Training epochs	1
Loss Weights	
λ_{grpo}	1.0
λ_{distill}	1.0
Policy clip ratio ϵ	0.2
IS clip threshold c	2.0
Distillation	
Divergence	JSD ($\alpha = 0.5$)
Top-k	100
$\tau_{\text{low}}, \tau_{\text{high}}$	0.3, 0.5
Reward Components	
Format bonus $\alpha_{\text{think}}, \alpha_{\text{ans}}$	0.3, 0.3
Length penalty $\tau_{\text{len}}, \gamma_{\text{len}}, \delta_{\text{len}}$	512, 3×10^{-4}, 1.0
Repetition penalty $\tau_{\text{rep}}, \beta_{\text{rep}}$	0.4, 1.0
n-gram size, tail chars	10, 100
Sequence Lengths	
Max prompt length	3072
Max response length	1024
Infrastructure	
GPUs	2
Sharding	FSDP2 (strategy 2)
vLLM GPU utilization	0.85
EMA & SFT	
EMA teacher decay τ_{ema}	0.05
SFT data	Qwen3.6-Plus trajectories

Table 5: Full hyperparameter configuration.

C Implementation Details

C.1 Model Architecture

We use Qwen3.5-2B as the base model, a vision-language model with 2B parameters supporting multi-image inputs. The model processes images through a vision encoder and generates text responses autoregressively through a language decoder.

C.2 Training Infrastructure

We train on 2 NVIDIA GPUs using FSDP2 with full sharding (sharding strategy 2). We use vLLM (Kwon et al., 2023) for efficient rollout generation with GPU memory utilization of 0.85. Gradient checkpointing and activation offloading are enabled for memory efficiency. We use Flash Attention 2 (Dao, 2024) for the attention implementation. Dynamic batch sizing is enabled with a maximum of 16,384 tokens per GPU for actor, rollout log probability, and reference log probability computations.

C.3 Teacher Input Construction

The teacher receives the same prompt as the student but with annotated images replacing the originals: $x_{\text{tea}} = (p, \tilde{I}_1, \tilde{I}_2)$, where $\tilde{I}_1$ and $\tilde{I}_2$ overlay bounding boxes highlighting the difference region on the original and edited images respectively. The student receives only the raw images: $x_{\text{stu}} = (p, I_1, I_2)$.

D Dataset Construction Details

Our dataset generation pipeline consists of five stages.

Stage 1: Taxonomy Construction. We construct a hierarchical object taxonomy from WordNet (Fellbaum, 1998), spanning 25 root categories (mammals, birds, insects, marine life, vehicles, buildings, furniture, appliances, tools, instruments, clothing, food, etc.) with adaptive depth pruning (max depth 7). Non-visual and abstract concepts are filtered via a blocklist. This taxonomy provides seed concepts for diverse scene generation.

Stage 2: Scene Generation. A large language model (Qwen3-14B) generates scene descriptions given 5 randomly sampled seed concepts from the taxonomy. Each scene specifies a setting theme, a detailed text-to-image prompt, and a list of 10–15 distinct objects. Scenes are designed in a "spot the difference" game style with dense, cluttered layouts. We generate plans for 2,000 scenes.

Stage 3: Image Generation. A text-to-image diffusion model (Z-Image-Turbo) renders original images at 1344×768 resolution. Each image is randomly assigned one of 8 artistic styles: flat vector art, 3D Pixar cartoon, anime, watercolor, comic book, claymation, isometric pixel art, or paper cutout craft. A negative prompt discourages photorealistic outputs.

Stage 4: Object Detection and Edit Planning. Grounding DINO (Liu et al., 2024) detects candidate objects and their bounding boxes in each generated image using the object list from Stage 2 as detection prompts (box threshold 0.45, text threshold 0.40). Objects whose area falls outside 0.5%–50% of the image area are excluded. The LLM then selects 1–3 objects per image and plans edits of three types: *replacement*, *color change*, or *removal*, generating precise instructions for an image editing model.

Stage 5: Image Editing. An image editing model (Qwen-Image-Edit) applies each planned edit. The target region is cropped, edited, and seamlessly reintegrated into the original image via Poisson blending (Pérez et al., 2003). The first planned edit per image is used to create the final image pair. Trace images—which overlay the edit bounding box on both the original and edited images—are generated as the teacher's privileged input.

D.1 Dataset Quality

We assess the quality of our synthetic dataset along several dimensions. **Edit fidelity:** By construction, the pipeline guarantees that a selected region is modified: the editing model receives a cropped region with explicit modification instructions, and Poisson blending reintegrates the result. Since the task requires only *localizing* the changed region—not interpreting its semantic content—the ground-truth bounding box is valid as long as the edit produces any visible change within the target area, which the pipeline ensures by design. Manual inspection of randomly sampled pairs confirmed that all edits were well-localized with accurate ground-truth boxes. **Box accuracy:** Ground-truth boxes come directly from the Grounding DINO detection stage (before editing), so they are spatially aligned with the edited region by construction. **Edit-type distribution:** The final 1,142 pairs include replacement (48%), color change (31%), and removal (21%) edits. **Box-size distribution:** Bounding boxes span 0.5%–50% of the image area by design (enforced by the area filter in Stage 4), with a median of 8.3%. **Filtering:** Of 2,000 generated scenes, 1,142 passed all quality filters (successful rendering, successful detection, successful editing, and area constraints); the remaining 858 were discarded due to detection failures (no objects matched), editing failures (inpainting artifacts), or area violations.

D.2 Prompt Template

The following prompt template is used for both training and evaluation:

```
You are given two images. Find the single
difference between them and provide a
bounding box around the changed region.

Think step by step in <thinking> tags,
then provide your answer in <answer> tags
as [x1, y1, x2, y2] where coordinates
are in [0, 1000] scale.
```

The teacher receives the same prompt but with annotated images $(\tilde{I}_1, \tilde{I}_2)$ replacing the raw images.

D.3 Data Format

Each sample is stored as a JSONL entry containing: image paths for original and edited images, ground-truth bounding box coordinates in $[x_1, y_1, x_2, y_2]$ format, annotated image paths for teacher input, and image dimensions (1344×768). Bounding boxes are provided in absolute pixel coordinates. The v2 format additionally normalizes coordinates to a $[0, 1000]$ scale and includes structured chain-of-thought prompting.

D.4 Bounding Box Parsing

Model responses are parsed by searching for an `<answer>` tag containing four comma-separated integers $[x_1, y_1, x_2, y_2]$ in $[0, 1000]$-normalized coordinates. If the tag is absent or the integers cannot be extracted, the prediction is marked as a parsing failure (IoU $= 0$). Predicted coordinates are clipped to $[0, 1000]$ and converted to absolute pixel coordinates for IoU computation against the ground truth.

E Additional Results

E.1 Divergence Type and Distillation Mode

E.2 Active Token Fraction

E.3 Additional Dataset Examples

Figure 7 shows two additional examples from our synthetic spot-the-difference dataset, illustrating the diversity of scenes and edit types.

Configuration	IoU@0.5 ↑	Mean IoU ↑
VisionRL-SD (JSD + Top-k)	**63.7**	**0.5928**
Divergence type		
Forward KL (α=0.0)	62.9	0.5846
Reverse KL (α=1.0)	63.3	0.5893
Distillation mode		
Full Logit	62.7	0.5887

Table 6: Divergence type and distillation mode comparison. JSD outperforms both forward and reverse KL, though the spread is small (0.8 points). Top-k outperforms full logit matching by 1.0 point.

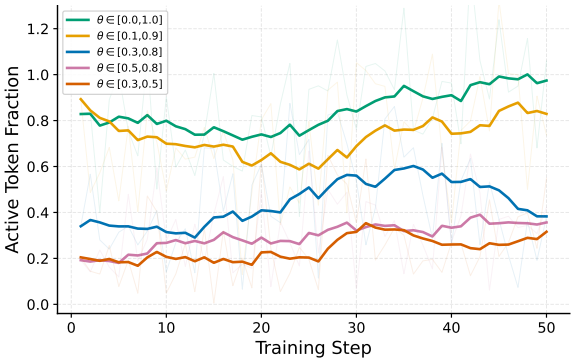


Figure 6: Active token fraction over training for different band-pass thresholds $[\tau_{\text{low}}, \tau_{\text{high}}]$.

E.4 RefCOCO Visual Grounding Evaluation

To assess generalization beyond the spot-the-difference task, we evaluate untrained Qwen3.5-2B and Qwen3.5-4B on the standard RefCOCO visual grounding benchmark (Kazemzadeh et al., 2014; Yu et al., 2016) (Table 7). The model is prompted to localize a referred object and output a bounding box in $[0, 1000]$-normalized coordinates.

Model	Mean IoU	Acc@0.5
Qwen3.5-2B	0.157	13.1
Qwen3.5-4B	0.152	13.3
Qwen3.5-2B + VisionRL-SD	0.167	13.7
Qwen3.5-4B + VisionRL-SD	0.164	14.3

Table 7: RefCOCO visual grounding results (val split). Training on our spot-the-difference task transfer to RefCOCO grounding a little.

Figure 7: Additional dataset examples. Top row: student inputs (raw original and edited images). Bottom row: teacher inputs (annotated with bounding boxes highlighting the difference region).