PIXEL-OPSD: BEYOND SCALAR REWARDS FOR PIXEL-LEVEL MLLMS VIA ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pixel-level multimodal large language models must preserve a region's meaning in language and its extent in pixels. Scalar outcome rewards assess the final prediction but do not specify how to correct its semantic and spatial errors. We introduce PIXEL-OPSD, which shifts pixel-level learning from scalar-reward reinforcement learning to on-policy self-distillation (OPSD). Our central idea is to make evidence-conditioned self-teaching responsive to both the semantic state of a generated trajectory and the spatial effects of its mask codes. Reliability routing adapts semantic correction to the sampled description through verified replacement, on-policy prefix refinement, or retention. Evidence-Guided Credit Assignment (EGCA) grounds mask-code credit in changes between decoded partial masks, identifying contributions that a terminal score obscures. These two mechanisms make OPSD sensitive to how a trajectory should be corrected and where spatial credit should be assigned. Supervised mixing supports this self-teaching process by anchoring grounding and descriptive completeness to human annotations. Experiments demonstrate consistent improvements over the baseline across region understanding and localization. PIXEL-OPSD thus advances pixel-level learning beyond scalar outcome feedback through trajectory-adaptive semantic correction and spatially localized credit, with all teaching mechanisms confined to training.

1 INTRODUCTION

Pixel-level multimodal large language models (MLLMs) connect language to specific image regions, enabling both region description and localization (Lai et al., 2024; Ren et al., 2024; Rasheed et al., 2024). Discrete mask tokens bring spatial outputs into the same autoregressive interface as language (Zhou et al., 2026). Yet a shared interface does not make their learning requirements identical. A description must preserve the target's identity and distinguishing properties, whereas a mask must capture its spatial extent. Successful pixel-level generation therefore requires semantic and spatial precision together: an accurate boundary cannot rescue a description of the wrong instance, and an informative description cannot compensate for a misplaced mask.

Recent reinforcement-learning approaches couple these capabilities through outcome feedback. CycleGRPO (Zhang et al., 2026), for example, evaluates whether a generated description allows the target region to be reconstructed and uses this score for policy optimization (Schulman et al., 2017). The resulting scalar assesses the completed cycle without explicitly specifying how to correct it. A wrong instance name, a missing attribute, and an inaccurate boundary may all lower reconstruction quality, but require different interventions. Correcting the first may require abandoning the current description; correcting the second may benefit from retaining its useful context; correcting the third requires identifying the mask codes responsible for the spatial change. The challenge is thus to provide feedback that addresses both the state of the semantic trajectory and the spatial effects of the generated codes.

We address this challenge through *on-policy self-distillation* (OPSD). On-policy distillation queries a teacher at student-visited states, providing distributional feedback on the student's own generations (Agarwal et al., 2024). In our setting, an evidence-conditioned self-teacher uses training-time

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.

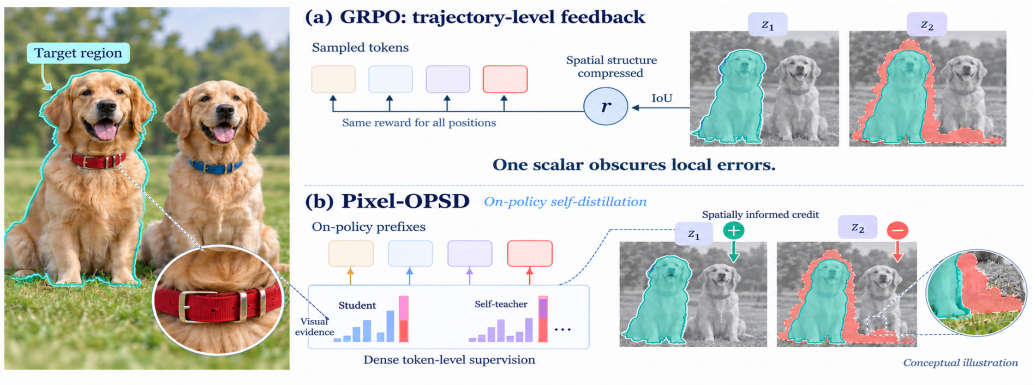


Figure 1: The overview separates the scalar-credit failure from our intervention. A trajectory-level cycle score aliases noun, relation, and boundary errors; Pixel-OPSD supplies token-level teaching and localizes spatial credit to decoded mask codes.

region information to guide learning from the model's sampled trajectories. The central contribution of PIXEL-OPSD is this shift in the learning framework, from scalar-reward reinforcement learning to OPSD for pixel-level understanding and localization. CycleGRPO is a separate RL baseline; its caption-to-localization cycle provides a point of comparison, rather than the GRPO implementation underlying our method. Our guiding principle is to make this self-teaching *adaptive to the semantic state of a trajectory and grounded in the spatial effects of its mask codes*. Reliability routing and spatial credit assignment realize these complementary requirements within a single OPSD framework.

Semantic correction must adapt to the trajectory. When a sampled description has entered an erroneous semantic path, conditioning further teaching on that prefix can constrain the correction itself. A recoverable description, however, can benefit from local guidance without discarding useful context. We therefore use reconstruction reliability to select the form of correction: verified self-teacher replacement for failed descriptions, on-policy prefix refinement for recoverable ones, and retention for reliable ones. Routing makes the correction depend on the sampled trajectory, directing self-teaching toward semantic errors while preserving useful generations.

Spatial credit must reflect the effect of each mask code. Adapting caption correction does not identify which code improves or damages the decoded region. We introduce *Evidence-Guided Credit Assignment* (EGCA) to connect code-level learning to changes between consecutive partial-mask reconstructions. By combining observed quality increments with an evidence-conditioned estimator, EGCA assigns spatial credit according to a code's contribution to the evolving mask. Routing determines how to correct the semantic trajectory; EGCA identifies where to assign credit in the spatial output. Together, they make the OPSD framework responsive to the two forms of precision required by pixel-level generation.

This self-teaching process also needs to preserve information that cycle consistency alone does not require. A description can allow the model to reconstruct a region while omitting attributes, parts, or relations important to a human reader. We therefore support OPSD with supervised mixing that anchors grounding and descriptive completeness to human annotations. These anchors keep self-teaching aligned with human referring expressions and informative descriptions. The self-teacher, evidence, and correction mechanisms are used only during training, preserving the model's inference interface.

We evaluate region localization and understanding on RefCOCO/+/g, GroundingSuite, and DLC-Bench. Comparisons with scalar-reward RL and ablations within the OPSD family show consistent improvements over the baseline and examine the roles of spatial credit and supervised mixing. Our contributions are:

- We introduce PIXEL-OPSD, an OPSD framework that moves pixel-level learning beyond scalar-reward supervision through evidence-conditioned self-teaching adapted to semantic trajectories and spatial outputs.

- We develop reliability routing and EGCA as complementary mechanisms for trajectory-adaptive semantic correction and spatially localized mask-code credit, supported by supervised mixing for human-grounded semantic coverage.
- We evaluate PIXEL-OPSD across region localization and understanding benchmarks, comparing OPSD with scalar-reward RL and examining its components with a shared set of metrics.

2 RELATED WORK

Pixel-level multimodal models. Region-aware MLLMs connect language to explicit masks through segmentation modules or discrete spatial representations (Lai et al., 2024; Ren et al., 2024; Rasheed et al., 2024; You et al., 2024; Yuan et al., 2024). SAMTok provides an autoregressive mask interface (Zhou et al., 2026), while region-understanding systems expand the semantic detail associated with localized inputs (Wang et al., 2026). Our focus is the learning signal for a pixel-level actor, rather than a new mask representation or inference architecture.

Related directions also include unified segmentation with OMG-Seg (Li et al., 2024), multi-purpose segmentation with RMP-SAM (Xu et al., 2025), and multimodal diffusion modeling with MMaDA (Tian et al., 2026). These works provide broader context for the segmentation and generation capabilities studied here.

Scalar-reward reinforcement learning. CycleGRPO couples region description and localization through reconstruction-based rewards (Zhang et al., 2026). Such feedback evaluates task outcomes, but does not directly provide a target distribution at each generated token. We study OPSD as a distinct approach to learning these capabilities. Our use of reconstruction evidence is directed toward constructing and routing self-teaching feedback for student-generated trajectories.

Preference-based alignment provides another perspective on the learning signal. Direct Preference Optimization (DPO) (Rafailov et al., 2023) learns from response preferences, whereas the distillation formulation considered here uses teacher feedback on student-generated trajectories.

On-policy distillation and self-teaching. On-policy distillation trains on student-generated histories, addressing the mismatch between teacher-generated training sequences and states visited by the student (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024). PIXEL-OPSD instantiates on-policy self-distillation with evidence-conditioned self-teacher feedback for pixel-level outputs. Its routing handles the recoverability of sampled descriptions, and EGCA connects the correction signal to decoded spatial changes. Supervised mixing complements self-teaching with human-grounded information that reconstruction consistency alone need not preserve (Wang et al., 2024; Liu et al., 2023a;b; Ouyang et al., 2022).

3 METHOD

3.1 PRELIMINARY

Problem setup. Given an image I and target region $M^\star$, we optimize a shared pixel-level actor π_θ for two coupled capabilities: describing the specified region and localizing the region referred to by a description. We denote an actor-generated caption by C_i and its corresponding localization rollouts by $\mathbf{Z}_i = \{Z_{i,k}\}_{k \in \mathcal{K}_i}$, where i indexes captions and k indexes localization responses conditioned on C_i.

CycleGRPO. As shown in Figure 2, CycleGRPO connects these capabilities in a caption-to-localization cycle. The actor first describes $(I, M^\star)$ with C_i and then predicts $Z_{i,k}$ from (I, C_i). A frozen decoder D maps each response to a pixel mask, and an evaluator $\mathcal{Q}$ provides two rewards: the rollout reward $s_{i,k}$ and its caption-level cycle average R_i^{cyc}:

$$s_{i,k} = \mathcal{Q}(D(Z_{i,k}), M^\star), \qquad R_i^{\text{cyc}} = \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} s_{i,k}. \tag{1}$$

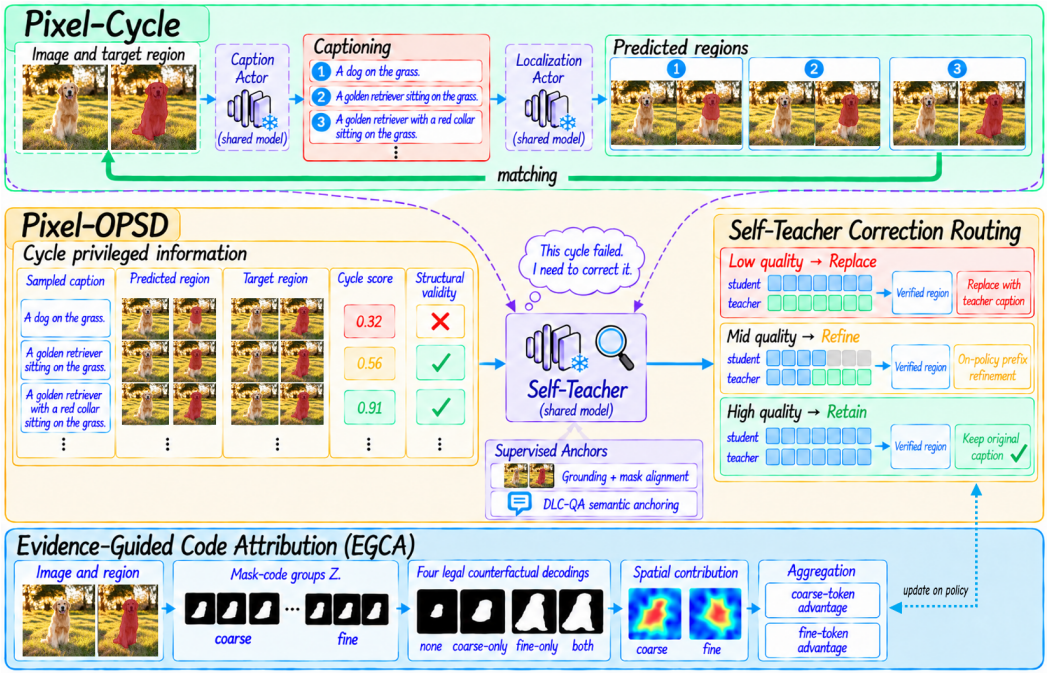


Figure 2: Overview of PIXEL-OPSD. Previous RL method provides trajectory-level rewards for caption-to-localization rollouts. Reliability-routed OPD refines caption tokens, EGCA assigns spatial credit to mask codes, and supervised anchors strengthen spatial and semantic coverage. All signals update a shared actor without changing the inference interface.

CycleGRPO group-normalizes these rewards into the caption advantage A_i^{cap} and localization advantage $A_{i,k}^{\text{loc}}$ and applies its standard clipped policy functional $\mathfrak{P}_\theta$. This scalar feedback closes the training cycle but assigns uniform credit to every token in a trajectory, obscuring whether an error originates from the description or a particular mask code. PIXEL-OPSD addresses this limitation with token-level caption correction, evidence-guided mask-code attribution, and supervised anchors.

3.2 RELIABILITY-ROUTED CAPTION CORRECTION

On-Policy Distillation (OPD). Following OPD (Agarwal et al., 2024), we replace uniform caption credit with token-level guidance on states visited by the current actor. The student samples the caption and localization path

$$C_i \sim \pi_\theta(\cdot \mid I, M^\star), \qquad Z_{i,k} \sim \pi_\theta(\cdot \mid I, C_i), \qquad \tau_i = (C_i, \mathbf{Z}_i) \sim \Pi_\theta(\cdot \mid I, M^\star). \tag{2}$$

A frozen teacher diagnoses and evaluates each visited caption prefix $h_{i,t}$ by utilizing privileged evidence $\mathcal{E}_i$. Let $p = p_{i,t} = \pi_\theta(\cdot \mid h_{i,t})$ and $q = q_{i,t} = \pi_\phi(\cdot \mid h_{i,t}, \mathcal{E}_i)$ be the student and teacher distributions over language tokens, their discrepancy is

$$\begin{aligned} \mathrm{JSD}_\beta(p, q) &= (1-\beta) D_{\mathrm{KL}}(p \| m) + \beta D_{\mathrm{KL}}(q \| m), \quad 0 < \beta < 1, \\ m &= (1-\beta)p + \beta q \end{aligned} \tag{3}$$

Each language position is weighted by normalized teacher confidence $w_{i,t}^{\text{tok}} \propto \exp[-H(q_{i,t})]$. OPD therefore supplies fine-grained correction without changing the on-policy sampling path. We next use cycle reliability to determine when such prefix correction is appropriate.

The cycle score measures whether a caption retains sufficient information to recover the target region, while $v_i \in \{0, 1\}$ indicates its structural validity. We use these two signals to select the least invasive correction warranted by the evidence:

$$\kappa_i = \begin{cases} \text{replace}, & v_i = 0 \text{ or } R_i^{\text{cyc}} < \tau_{\text{low}}, \\ \text{refine}, & v_i = 1,\ \tau_{\text{low}} \le R_i^{\text{cyc}} < \tau_{\text{high}}, \\ \text{retain}, & v_i = 1,\ R_i^{\text{cyc}} \ge \tau_{\text{high}}. \end{cases} \tag{4}$$

The routing mode κ_i specifies how the sampled caption is treated, rather than selecting among different base objectives. The ordered thresholds satisfy $\tau_{\text{low}} < \tau_{\text{high}}$. Low-reliability or structurally invalid captions are replaced by verified targets, intermediate captions are refined on their sampled prefixes, and high-reliability captions are retained without auxiliary correction. The advantage remains active in all three cases when the caption is structurally valid. An invalid caption receives no caption policy update and can be corrected only through verified replacement.

Verified replacement for failed captions. Prefix distillation is inappropriate when a caption omits or misidentifies the target. Such a trajectory may be wrong almost in its entirety, in which case even a privileged teacher has limited ability to recover while remaining conditioned on the erroneous prefix. For $\kappa_i = \text{replace}$, the teacher instead uses privileged evidence to sample a set of fresh caption candidates:

$$\widehat{C}_i^{(m)} \sim \pi_\phi(\cdot \mid I, M^\star, \mathcal{E}_i), \qquad m = 1, \dots, M. \tag{5}$$

The actor independently reruns localization for each candidate, and the resulting reconstruction is assessed by the same decoder and evaluator used in the cycle:

$$\begin{aligned} &\widehat{Z}_{i,k}^{(m)} \sim \pi_\theta(\cdot \mid I, \widehat{C}_i^{(m)}), \\ &\widehat{R}_i^{(m)} = \frac{1}{K_{\text{ver}}} \sum_{k=1}^{K_{\text{ver}}} \mathcal{Q}\Big(D(\widehat{Z}_{i,k}^{(m)}), M^\star\Big), \\ &m_i^\star = \arg\max_m \widehat{R}_i^{(m)}, \qquad a_i^{\text{ver}} = \mathbf{1}\Big\{v(\widehat{C}_i^{(m_i^\star)}) = 1,\ \widehat{R}_i^{(m_i^\star)} \geq \tau_{\text{ver}}\Big\}. \end{aligned} \tag{6}$$

Only a structurally valid candidate that passes the cycle threshold can become the replacement target $\widetilde{C}_i = \widehat{C}_i^{(m_i^\star)}$. Let $\mathfrak{T}_\theta(Y \mid x)$ denote the actor's token-averaged teacher-forcing negative log-likelihood for target sequence Y given x. The replacement confidence is determined by its verified improvement over the actor caption:

$$w_i^{\text{rep}} = \text{clip}\left(\frac{\widehat{R}_i^{(m_i^\star)} - R_i^{\text{cyc}}}{1 - R_i^{\text{cyc}}}, 0, 1\right), \qquad \mathcal{C}_i^{\text{rep}} = a_i^{\text{ver}} w_i^{\text{rep}} \mathfrak{T}_\theta(\widetilde{C}_i \mid I, M^\star). \tag{7}$$

If no candidate passes verification, no replacement loss is applied, and the original valid localization rollouts contribute only through CycleGRPO and EGCA. Regeneration targets are therefore selected by recoverability under the task cycle, not by the teacher's linguistic fluency alone.

On-policy prefix refinement for recoverable captions. An intermediate-reliability caption has already entered a useful semantic path but may still contain an ambiguous noun, omit a distinguishing attribute, or terminate prematurely. The teacher and actor therefore score the same actor-generated history $h_{i,t}$, while only the teacher observes $\mathcal{E}_i$. We use a sample-level routing weight that decreases as the caption approaches the retain boundary:

$$w_i^{\text{route}} = \text{clip}\left(\frac{\tau_{\text{high}} - R_i^{\text{cyc}}}{\tau_{\text{high}} - \tau_{\text{low}}}, w_{\min}, 1\right), \qquad \mathcal{C}_i^{\text{ref}} = \frac{a_i^{\text{conf}} w_i^{\text{route}}}{|\mathcal{T}_i^{\text{lang}}|} \sum_{t \in \mathcal{T}_i^{\text{lang}}} w_{i,t}^{\text{tok}} \, \text{JSD}_\beta(p_{i,t}, q_{i,t}), \tag{8}$$

Here $a_i^{\text{conf}} = \mathbf{1}\{R_i^{\text{cyc}} \geq \tau_{\text{conf}}\}$ is a cycle-evidence gate that filters intermediate captions unsupported by sufficiently reliable reconstructions. Restricting the sum to $\mathcal{T}_i^{\text{lang}}$ prevents privileged geometry from leaking into mask-code or object-reference tokens. Because these histories are sampled from π_θ, the teacher corrects states the actor will encounter at inference rather than forcing it to imitate an off-policy reference sequence. The complete caption correction is expressed as a single routed operator:

$$\mathcal{C}_i(\theta) = \begin{cases} \mathcal{C}_i^{\text{rep}}, & \kappa_i = \text{replace}, \\ \mathcal{C}_i^{\text{ref}}, & \kappa_i = \text{refine}, \\ 0, & \kappa_i = \text{retain}. \end{cases} \tag{9}$$

The three branches are mutually exclusive. Averaging $\mathcal{C}_i$ over the complete cycle batch naturally accounts for the number of active replacement and refinement examples, avoiding separate coefficients for regeneration and distillation.

3.3 Evidence-Guided Code Attribution

Routing corrects captions but cannot determine whether a spatial error arose from a coarse code or its fine refinement. Let the ℓ-th valid mask group be $G_{i,k,\ell} = (c_{i,k,\ell}, f_{i,k,\ell})$ and $P_{i,k,\ell-1}$ its preceding sampled groups. Because the decoder requires a complete pair, we use fixed reference codes $(c_{\mathrm{r}}, f_{\mathrm{r}})$ to construct four legal counterfactuals $S \in \{\varnothing, (c, f_{\mathrm{r}}), (c_{\mathrm{r}}, f), (c, f)\}$ and evaluate $v^S = \mathcal{Q}(D(P_{i,k,\ell-1} \oplus S), M^\star)$. The coarse and fine contributions are their Shapley values:

$$\phi^c_{i,k,\ell} = \tfrac{1}{2}[(v^c - v^0) + (v^{cf} - v^f)], \quad \phi^f_{i,k,\ell} = \tfrac{1}{2}[(v^f - v^0) + (v^{cf} - v^c)], \tag{10}$$

so $\phi^c + \phi^f = v^{cf} - v^0$. Suppressing the shared (i, k, ℓ) indices below, a learnable evidence head predicts each contribution from frozen EMA-teacher evidence and the counterfactual mask summaries, and is calibrated only against detached Shapley targets:

$$\begin{aligned} e^d &= \mathcal{H}_\psi(F_\phi(\mathcal{E}_i), \mathcal{V}_{i,k,\ell}, d, R_i^{\mathrm{cyc}}), \\ \mathcal{L}_{\mathrm{cal}} &= \mathbb{E}\left[\tfrac{1}{2} \sum_{d \in \{c,f\}} (e^d - \mathrm{sg}(\phi^d))^2\right]. \end{aligned} \tag{11}$$

where $\mathcal{V}_{i,k,\ell}$ summarizes the four decoded counterfactuals. We blend observed and predicted evidence, project the pair to preserve the group contribution, and add the resulting bounded credit only to its corresponding token:

$$\begin{aligned} u^d &= \alpha\, \mathrm{sg}(\phi^d) + (1 - \alpha)\, \mathrm{sg}(e^d), \\ \bar{g}^d &= u^d - \tfrac{1}{2}[u^c + u^f - \mathrm{sg}(v^{cf} - v^0)], \quad g^d = \mathrm{clip}(\bar{g}^d, -g_{\max}, g_{\max}), \\ \widetilde{A}_{i,k,t} &= A^{\mathrm{loc}}_{i,k} + \lambda_{\mathrm{e}} \sum_{\ell,d} \mathbf{1}\{t = t^d_{i,k,\ell}\} g^d_{i,k,\ell}. \end{aligned} \tag{12}$$

EGCA therefore assigns distinct credit to coarse selection and fine boundary decisions without changing the cycle reward or adding another actor loss. No-target rollouts bypass EGCA, and the decoder, privileged evidence, and evidence head are training-only.

3.4 Supervised Anchors

The self-generated cycle may admit a communication shortcut: captions remain recoverable by the model's locator while omitting human-relevant semantics, and homogeneous rollouts provide weak supervision for reasoning-based segmentation. We therefore introduce two supervised streams that anchor reasoning segmentation, mask alignment, and semantic completeness without changing the inference model.

Reasoning segmentation and mask alignment. For $(I_j, Q_j, M^\star_j, Y_j) \sim \mathcal{D}_{\mathrm{ref}}$, we apply standard mask-alignment supervision to the sampled response $Z_{j,k} \sim \pi_\theta(\cdot \mid I_j, Q_j)$. Its group-normalized segmentation score is optimized through the shared policy objective, while teacher forcing on Y_j provides a stable supervised anchor.

DLC-QA caption anchoring. We construct a dataset $\mathcal{D}_{\mathrm{qa}}$ containing n DLC-QA examples. Each example consists of an image I_j, a target region $M^\star_j$, and a set of multiple-choice questions, candidate options, and answer keys derived from human-annotated detailed localized captions (DLCs):

$$\mathbb{Q}_j = \{(q_{j,r}, \mathcal{O}_{j,r}, y_{j,r})\}_{r=1}^{N_j}. \tag{13}$$

During training, the actor independently samples a group of student descriptions $\mathbf{C}_j \sim \pi_\theta(\cdot \mid I_j, M^\star_j)$. A fixed text-only evaluator answers each question using only a sampled description and the candidate options:

$$s^{\mathrm{qa}}_{j,k} = \frac{1}{N_j} \sum_{r=1}^{N_j} \mathbf{1}\{\mathcal{A}_{\mathrm{qa}}(C_{j,k}, q_{j,r}, \mathcal{O}_{j,r}) = y_{j,r}\}, \qquad b_{j,k} = \mathcal{N}_{\mathcal{G}_j}(s^{\mathrm{qa}}_{j,k}). \tag{14}$$

Answer keys are used only to compute accuracy; the evaluator does not observe the target mask or the human-authored DLC. The resulting group-relative advantage thus measures how much verifiable semantic information each student description preserves, rather than its lexical similarity to a particular reference. The pair $(C_{j,k}, b_{j,k})$ enters the shared policy functional directly, so DLC-QA changes the reward and data source without introducing a new loss family.

3.5 Joint Training

Let $\mathcal{S} = \{\mathrm{C}, \mathrm{L}, \mathrm{G}, \mathrm{Q}\}$ denote cycle captioning, cycle localization, supervised grounding, and DLC-QA. Each source provides response X_s and group-normalized advantage $\mathbf{A}_s$ to the same policy functional; only localization advantages receive EGCA credit. The complete actor objective is

$$\begin{aligned}\mathcal{L}_{\text{actor}}(\theta) = &\sum_{s\in\mathcal{S}} \omega_s \mathbb{E}_{\mathcal{D}_s,\pi_\theta}[\,\mathfrak{P}_\theta(X_s, \widetilde{\mathbf{A}}_s)\,] \\ &+ \omega_{\mathrm{C}} \mathbb{E}_{\mathcal{D}_{\mathrm{C}}}[\,\mathcal{C}_i(\theta)\,] + \lambda_{\text{align}} \mathcal{L}_{\text{align}},\end{aligned} \tag{15}$$

where $\widetilde{\mathbf{A}}_s$ equals $\mathbf{A}_s$ except at the coarse and fine token positions receiving EGCA credit. The actor is optimized only through the shared policy functional, while ψ is optimized separately by $\mathcal{L}_{\text{cal}}$; all Shapley and evidence credits are detached before entering $\widetilde{\mathbf{A}}_s$. Removing routed correction, EGCA, and exact alignment recovers the shared GRPO objective. The teacher, target-conditioned counterfactual decoding, and EGCA head are training-only, so inference is unchanged.

4 Experiments

4.1 Experimental Setup

We use SAMToK-4B (Zhou et al., 2026) as our baseline. Self-supervised training uses 20k COCO images with masks, same as the training-data scale of CycleGRPO (Zhang et al., 2026). For supervised anchor, we additionally use 10k examples drawn from the RefCOCO and DAM (Lian et al., 2025). We train on eight NVIDIA H20 GPUs; further experimental details are provided in the Appendix.

We report validation cIoU on RefCOCO/+/g and overall cIoU on MR-RefCOCO/+/g and MR-PACO. We evaluate general grounding with GroundingSuite gIoU (Hu et al., 2025) and region understanding with DLC-Bench positive, negative, and average accuracy.

4.2 Main Results

We compare Pixel-OPSD with pixel-level models across localization and region understanding (Tables 1–2; Figure 3). Against CycleGRPO, our RefCOCO/+/g scores of 81.1/75.5/77.7 improve by 4.6/6.8/2.7 points. We also lead the reported multi-round comparisons, reaching 51.5 on MR-PACO. DLC-Bench average accuracy rises from 65.0 to 68.8, and GroundingSuite All from 68.2 to 68.8. Grounding gains concentrate in Stuff and Multi (+1.4/+2.8); Part is unchanged and Single declines by 1.0. Relative to SAMToK-4B, we gain 11.9 points on GroundingSuite All and 7.0 on MR-PACO, supporting improvements in both grounding and interactive localization.

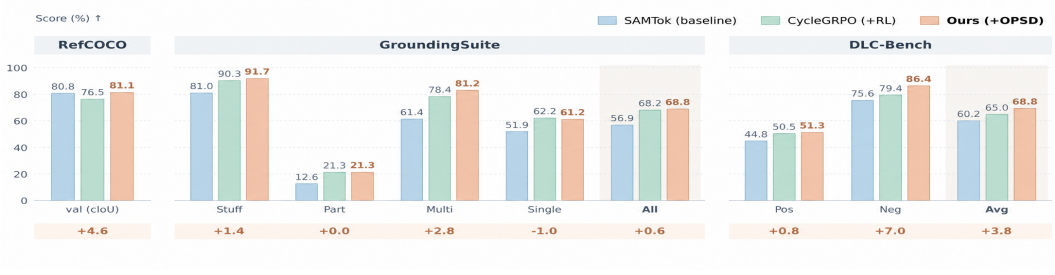


Figure 3: **Benchmark-wise comparison of scalar-reward RL and OPSD.**

Table 1: **Referring and interactive segmentation.** RefCOCO/+/g: validation cIoU. MR benchmarks: overall cIoU. Baseline references for all tables are given in Appendix B.

Referring segmentation

Method	Size	RefCOCO	RefCOCO+	RefCOCOg
UniPixel	3B	80.6	74.4	76.2
InstructSeg	3B	80.0	71.1	75.5
SAMToK	4B	80.8	75.2	76.5
CycleGRPO	4B	76.5	68.7	75.0
PIXEL-OPSD	4B	**81.1**	**75.5**	**77.7**

Multi-round interactive segmentation

Method	Size	MR-RefCOCO/+/g			MR-PACO
		RefCOCO	RefCOCO+	RefCOCOg	
Sa2VA	4B	76.4	72.0	77.0	26.4
UniPixel	3B	73.8	68.5	75.4	37.6
InstructSeg	3B	53.2	49.5	60.3	24.1
SAMToK	4B	81.1	78.4	80.5	44.5
CycleGRPO	4B	79.1	76.0	77.5	50.0
PIXEL-OPSD	4B	**81.7**	**78.6**	**81.1**	**51.5**

Table 2: **Region captioning and general grounding.** DLC-Bench reports positive, negative, and average accuracy; GroundingSuite reports gIoU by subset and overall.

DLC-Bench

Method	Size	Pos.	Neg.	Avg.
GPT-4o	–	43.4	79.6	61.5
o1	–	46.3	78.8	62.5
DAM	3B	48.9	82.4	65.6
Sa2VA	4B	4.8	20.8	12.8
GAR	1B	48.9	**87.0**	67.9
SAMToK	4B	44.8	75.6	60.2
CycleGRPO	4B	50.5	79.4	65.0
PIXEL-OPSD	4B	**51.3**	86.4	**68.8**

GroundingSuite

Method	Size	Stuff	Part	Multi	Single	All
Sa2VA	4B	90.7	18.3	73.7	50.0	62.1
PaDT Pro	3B	28.3	10.3	44.8	32.3	31.1
UniPixel	3B	90.8	23.1	78.9	58.0	67.0
InstructSeg	3B	56.2	**24.9**	66.4	51.6	52.6
EVF-SAM	1B	83.9	13.1	66.9	37.9	53.3
SAMToK	4B	81.0	12.6	61.4	51.9	56.9
CycleGRPO	4B	90.3	21.3	78.4	**62.2**	68.2
PIXEL-OPSD	4B	**91.7**	21.3	**81.2**	61.2	**68.8**

Table 3: **Progressive component ablation.** RefCOCO: val cIoU; DLC-Bench: accuracy; GS: overall gIoU; MR-RefCOCO and MR-PACO: overall cIoU.

Method	RefCOCO	DLC-Bench			GS	MR-RefCOCO	MR-PACO
	val	Pos.	Neg.	Avg.	All	Overall	Overall
CycleGRPO	76.5	50.5	79.4	65.0	68.2	79.1	50.0
OPSD only	74.0	51.4	82.6	67.0	64.4	74.7	40.4
+ reliability routing	77.7	**52.7**	82.3	67.5	**69.0**	81.0	50.7
+ supervised mixing	80.9	52.2	84.6	68.4	68.6	81.6	51.1
+ EGCA (PIXEL-OPSD)	**81.1**	51.4	**86.3**	**68.8**	68.8	**81.7**	**51.5**

4.3 ABLATION STUDY

We conduct a progressive component ablation to assess reliability routing, supervised mixing, and EGCA (Table 3). OPSD alone improves DLC-Bench average accuracy to 67.0 but reduces localization scores. Reliability routing raises RefCOCO from 74.0 to 77.7 and MR-PACO from 40.4 to 50.7, showing the benefit of adapting semantic correction. Supervised mixing further improves these scores to 80.9 and 51.1, while GroundingSuite declines from 69.0 to 68.6. Adding EGCA yields 81.1 and 51.5, with DLC-Bench average accuracy reaching 68.8. The full model improves the balance across tasks, although routing alone retains the highest GroundingSuite score.

DLC-Bench average accuracy increases across the OPSD variants (67.0, 67.5, 68.4, and 68.8), while localization reveals component-specific trade-offs. These sequential additions do not isolate interactions, and supervised mixing introduces additional labeled data. More performance comparisons and parameter ablations are provided in Appendix C.

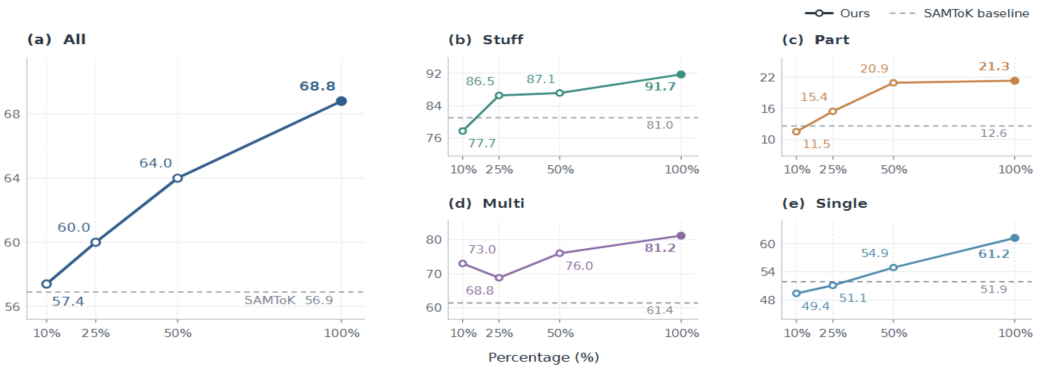


Figure 4: **Data scaling on GroundingSuite** (gIoU). Dashed lines denote baseline SAMTok.

4.4 Scaling with Training Data

We conduct a training-data ablation using 10%, 25%, 50%, and 100% of the data, as shown in Figure 4. Overall gIoU rises from 57.4 to 60.0, 64.0, and 68.8, gaining 11.4 points. Doubling from half to full data still yields 4.8 points. Part nearly saturates at 50% (20.9 versus 21.3), while Stuff and Single continue improving. Multi initially drops from 73.0 to 68.8 before recovering to 76.0 and 81.2. Overall scaling remains consistently positive despite these subset-specific differences.

4.5 Qualitative Comparison

We also compare masks for identical inputs and queries, as shown in Figure 5. Our IoUs of 0.86/0.93/0.96 exceed CycleGRPO's 0.08/0.66/0.27. Our predictions better select the cars, cover both trays, and delineate the two women while excluding the road sign, illustrating more precise target selection and spatial coverage.



Figure 5: **Matched mask predictions** for identical inputs and queries.

5 Conclusion

We presented Pixel-OPSD, an on-policy self-distillation framework that advances pixel-level multimodal learning beyond scalar outcome feedback. Its central principle is to adapt evidence-conditioned self-teaching to the semantic state of a generated trajectory and the spatial effects of its mask codes. Reliability routing selects how a description should be corrected, while EGCA grounds code-level credit in the evolution of the decoded mask. Supervised mixing supports this process by preserving human-grounded semantic coverage beyond what cycle consistency alone requires. Experiments show consistent improvements over the baseline across region understanding and localization. These results support an OPSD framework that couples trajectory-adaptive semantic correction with spatially localized credit while retaining the original inference interface.

AI USE STATEMENT

OpenAI ChatGPT/Codex was used to assist with manuscript editing, interpretation and presentation of author-provided experimental results. Language-model feedback within the research method, including self-teaching and question-based evaluation, is described in Section 3. The numerical results in the experimental tables and scaling curves were supplied by the authors; this writing and formatting assistance did not generate experimental measurements. The authors are responsible for the final manuscript, its scientific claims, and the validity of the reported results.

ETHICS STATEMENT

This work studies pixel-level vision-language understanding and localization using established benchmarks. Such models can inherit biases from their data and produce incorrect descriptions or spatial predictions. Benchmark performance does not establish reliability in safety-critical applications. Any downstream use should include independent validation and respect privacy, consent, and the licenses governing the underlying images and annotations.

REPRODUCIBILITY STATEMENT

Appendix A provides training pseudocode and implementation details, including the rollout record, data conversion, determinism checks, and logging schema. Appendix B documents benchmark reporting and baseline references; Appendix C gives detailed comparisons and the numerical data-fraction ablation. The teacher and evidence-related components are used only during training, preserving the original inference interface.

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 *International Conference on Learning Representations*, 2024.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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

Rui Hu, Lianghui Zhu, Yuxuan Zhang, Tianheng Cheng, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Groundingsuite: Measuring complex multi-granular pixel grounding. *arXiv preprint arXiv:2503.10596*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. In *International Conference on Machine Learning*, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Lin Li, Shilong Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

Xiangtai Li, Haobo Yuan, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, and Yin Cui. Describe anything: Detailed localized image and video captioning. In *IEEE/CVF International Conference on Computer Vision*, 2025.

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

Haotian Liu, Yuheng Li, Yifan Li, Bo Li, Yuanhan Zhang, et al. Improved baselines with visual instruction tuning. In *arXiv preprint arXiv:2310.03744*, 2023b.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Zhiliang Peng, Wenjun Wang, Li Dong, Yaru Hao, et al. Kosmos-2: Grounding multimodal large language models to the world. In *arXiv preprint arXiv:2306.14824*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Hanoona Rasheed, Muhammad Maaz, Hanoon Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Mubarak Khan. Glamm: Pixel grounding large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Nikhila Ravi, Valentin Gabeurion, Yuan-Ting Hu, Zhi Hu, et al. Sam 2: Segment anything in images and videos. In *arXiv preprint arXiv:2408.00714*, 2024.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26364–26373. IEEE, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. In *International Conference on Learning Representations*, 2024.

Ye Tian, Ling Yang, Jiongfan Yang, Anran Wang, Yu Tian, Jiani Zheng, Haochen Wang, Zhiyang Teng, Zhuochen Wang, Yinjie Wang, Yunhai Tong, Mengdi Wang, and Xiangtai Li. Mmada-parallel: Multimodal large diffusion language models for thinking-aware editing and generation. In *International Conference on Learning Representations*, 2026.

Haochen Wang, Yuhao Wang, Tao Zhang, Yikang Zhou, Yanwei Li, Jiacong Wang, Jiani Zheng, Ye Tian, Jiahao Meng, Zilong Huang, Guangcan Mai, Anran Wang, Yunhai Tong, Zhuochen Wang, Xiangtai Li, and Zhaoxiang Zhang. Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. In *International Conference on Learning Representations*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. In *arXiv preprint arXiv:2409.12191*, 2024.

Shilin Xu, Haobo Yuan, Qingyu Shi, Lu Qi, Jingbo Wang, Yibo Yang, Yining Li, Kai Chen, Yunhai Tong, Bernard Ghanem, et al. Rmp-sam: Towards real-time multi-purpose segment anything. In *International Conference on Learning Representations*, volume 2025, pp. 35567–35589, 2025.

Haoxuan You, Haotian Zhang, Zhe Gan, Xueqing Du, Bowen Zhang, Zhen Wang, Yu Cao, et al. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28202–28211. IEEE, 2024.

Xin Zhang, Haochen Wang, Yikang Zhou, Zhuochen Wang, Xiangtai Li, and Robby T. Tan. Actor as its own critic: Unifying region understanding and localization via cyclegrpo. In *European Conference on Computer Vision*, 2026.

Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, Haochen Wang, Haobo Yuan, Jiacong Wang, Lu Qi, Hao Fei, Shunping Ji, Anran Wang, Zhuochen Wang, Yujing Wang, Cheng Chen, and Xiangtai Li. Samtok: Representing any mask with two words. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xueyan Zou, Jiahui Yang, et al. Seem: Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems*, 2023.