FROM SPATIAL EVIDENCE TO TARGET UNDERSTANDING: ALIGNING DIFFUSION AND GROUNDING MODELS FOR RELIABLE VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual grounding maps an image and a referring query to a bounding box. Currently, text-to-image diffusion models appear well suited to augment this ability because of their dense visual features. However, we find that widely adopted diffusion augmentation can be harmful for visual grounding, which may perform even worse than noisy inputs. To diagnose this problem, we examine spatial evidence in model features, and find that naively generated images lead the model to disperse focus over the whole image, whereas noisy inputs coincidentally align with the grounding model's target-centered focus, yielding higher performance. This suggests that only aligned information from diffusion models benefits visual grounding. Motivated by this finding, we propose DIRECT, a diffusion-supervised framework that strengthens vision-grounding alignment from both the diffusion and grounding sides. It employs a frozen diffusion model as a teacher with a residual-feature mechanism to extract better-aligned features, and a grounding model as a student that uses paired valid and counterfactual queries to filter out misaligned diffusion-augmented features. The student then learns from the frozen teacher to capture detailed visual information, while only the VLM student is retained at inference. Extensive experiments on GroundingME, FineCops, gRefCOCO, LISA, and RefCOCOg show that DIRECT consistently improves diffusion-augmented visual grounding and outperforms state-of-the-art methods.

1 INTRODUCTION

Visual grounding maps an image and a referring query to a bounding box, or to `null` when the image does not support that query. Reliable grounding therefore requires both spatially precise localization and a judgement of whether all queried attributes and relations are valid (Kazemzadeh et al., 2014; Yu et al., 2016; He et al., 2023; Liu et al., 2024). Text-to-image diffusion models appear well suited to augment this ability: their large-scale text-to-visual pretraining may provide dense visual evidence that is absent from sparse box supervision (Zhao et al., 2023; Xu et al., 2023; Tang et al., 2023; Luo et al., 2023). Typically, the visual grounding model would take the visual features or samples as input for feature or data augmentation with consistent semantic relevance.

However, in this paper, we find that the widely adopted diffusion augmentation can be harmful for visual grounding: by taking both the diffusion-augmented samples and the noisy samples as input (Fig. 1a), we find that the diffusion-augmented performance is even lower than the noisy-input one (Fig. 1b). As in Fig. 1c, although the query-augmented image is semantically relevant to the target query image, these augmented images drive the model's prediction from the original ones (blue boxes) to worse ones (red boxes), with more shifts from the ground truth (green boxes). This experiment reveals a gap between semantic relevance and the visual grounding performance: what information is really needed to augment the visual grounding model?

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: Following current works, by taking both the diffusion-augmented samples or noisy samples into the visual grounding model (a), we find that the diffusion-augmented performance is even lower than the noise-input one (b). Although the semantic relevance in the diffusion-augmented samples is high in many test cases, including component, state, text, and relational descriptions, the predictions are still harmed by the diffusion models (c). In this paper, we aim to address this issue for better diffusion-augmented visual grounding.

To understand where this problem originates, we diagnose the diffusion-augmented model from the spatial evidence in its features. After carefully excluding irrelevant factors, we compare the ratio (the GT lift value) of within-bounding-box activation and the outside-bounding-box activation, where a higher value means more focus on the target object. We find that the naively generated image leads the model to have dispersed focus on the whole image, instead of focusing on the target object, leading to a low GT lift value. In contrast, for the noise-input model, as the noise tends to be applied to the center area of the image, it coincidentally aligns with the grounding model's focus on the target object, and leads to a higher GT lift value and performance. Inspired by this, we manually intervene in the diffusion process to explicitly strengthen the alignment of the diffusion and grounding model, and indeed observe an increase in the GT lift value. In all, we conclude that only the aligned information in the diffusion model can help the visual grounding task.

Based on the above analysis, for better diffusion-augmented visual grounding, we propose DIRECT, a diffusion-supervised framework to strengthen the vision-grounding alignment from both the diffusion side and the grounding side. Specifically, we take a frozen diffusion model as the teacher and design a residual-feature mechanism to extract features from the better-aligned diffusion model. Then, we take a grounding model as the student and pair valid/counterfactual queries for the same image for the grounding model to filter out misaligned diffusion-augmented features. Finally, the grounding model will learn from the frozen diffusion model to capture detailed visual information. At inference, only the VLM student is retained. In summary, our contributions are listed as follows.

- We find a misalignment problem between the diffusion and grounding models, making diffusion-generated samples even worse than pure noise when augmenting the grounding model.
- To address the misalignment problem for diffusion-augmented visual grounding, we propose DIRECT to take the diffusion model as the teacher and take the grounding model as the student, and design to strengthen the alignment from both the diffusion side and the grounding side.
- Extensive experiments on GroundingME, FineCops, gRefCOCO, LISA, and RefCOCOg show that we can consistently improve performance and outperform state-of-the-art works.

2 Related Work

Multimodal grounding and no-target recognition. Referring-expression comprehension traditionally assumes one valid referent and evaluates localization on ReferItGame and the RefCOCO family (Kazemzadeh et al., 2014; Mao et al., 2016; Yu et al., 2016). Architectures have progressed from proposal matching to end-to-end transformer grounding (Deng et al., 2021; Kamath et al., 2021; Liu et al., 2023) and instruction- tuned MLLMs that emit coordinates or masks (Chen et al.,

2023; Peng et al., 2024; You et al., 2024; Lai et al., 2024; Rasheed et al., 2024). This single-target assumption leaves behavior undefined when the expression is unsupported. gRefCOCO generalizes comprehension to multiple or zero targets (He et al., 2023); FineCops-Ref introduces controlled compositional positives and hard negatives (Liu et al., 2024); and GroundingME exposes appearance, spatial, capability-limited, and rejection failures in modern MLLMs (Li et al., 2026). We do not introduce another benchmark or a new output interface. Instead, we study how a generative teacher can improve localization without conflating spatial evidence with target existence, and train the same bbox-or-null student for both outcomes.

Diffusion representations for perception. Diffusion models learn rich intermediate representations while denoising images (Ho et al., 2020; Rombach et al., 2022; Esser et al., 2024). Their features support semantic correspondence and dense prediction without using generated pixels as labels: DIFT and Diffusion Hyperfeatures extract descriptors across layers or timesteps (Tang et al., 2023; Luo et al., 2023), while VPD and ODISE adapt text-to-image backbones to segmentation, depth, and open-vocabulary perception (Zhao et al., 2023; Xu et al., 2023). More recent work transfers generative knowledge into VLMs through reconstruction or semantic feedback. DS-VLM directly supervises visual representations with a diffusion decoder (Sun et al., 2025); DEEM aligns an image encoder using generative feedback (Luo et al., 2025); GenHancer uses denoising to strengthen vision-centric features (Ma et al., 2025); and VLV distills through a frozen diffusion decoder and a language bottleneck (Zhang et al., 2025). These works establish diffusion as a useful perceptual prior, but do not resolve the coordinate-frame mismatch and bbox-or-null ambiguity studied here. DIRECT keeps the teacher anchored to the input image, transfers regional relations rather than generated content, and treats rejection as a separate calibration problem.

Knowledge distillation for VLMs. Classical knowledge distillation matches softened outputs (Hinton et al., 2015); feature-based methods subsequently aligned intermediate activations or attention maps (Romero et al., 2015; Zagoruyko & Komodakis, 2017). Direct feature matching is fragile when teacher and student differ in architecture, resolution, and feature basis. Relational objectives instead preserve pairwise geometry or similarity structure (Park et al., 2019; Tung & Mori, 2019; Tian et al., 2020). This line of work motivates transferring spatial relations instead of requiring dimension-wise correspondence between heterogeneous teacher and student representations.

Representation reading and latent intervention. Representation engineering treats high-level activation directions as functional units that can be identified from contrastive stimuli and used to read or control model behavior (Zou et al., 2023). L2V-CoT extends this perspective across modalities: it extracts contrastive CoT directions from an LLM, retains and resamples their low-frequency components, and injects them into a VLM at inference time (Zhan et al., 2026). These studies show that contrastive activation structure can reveal task-relevant directions, but they address representation control and reasoning rather than paired localization and no-target grounding.

3 DIAGNOSING GENERATIVE PRIORS FOR GROUNDING

3.1 PRELIMINARIES

Given an image x and a query q, a visual grounding model f_θ predicts a bounding box or `null`:

$$f_\theta(x, q) \in \mathcal{B} \cup \{\texttt{null}\}. \tag{1}$$

We denote the ground-truth box by $b = (x_1, y_1, x_2, y_2)$. The VLM is trained autoregressively with a unified JSON answer. We denote this answer-token negative log-likelihood by $\mathcal{L}_{\mathrm{SFT}}$.

For the diffusion model, let E be a frozen VAE encoder and $D_\phi^{(\ell)}$ be the selected block of a frozen diffusion model (e.g., SD3.5 Transformer). We use the standard diffusion forward process to perturb the source-image latent and extract conditional and query-free features:

$$\begin{aligned} z_t &= \sqrt{\bar{\alpha}_t} E(x) + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \\ F &= D_\phi^{(\ell)}(z_t, t) \end{aligned} \tag{2}$$

The forward process follows standard diffusion models (Ho et al., 2020; Rombach et al., 2022), and the teacher follows the SD3 diffusion-transformer family (Esser et al., 2024).

Table 1: Controlled input comparison on GroundingME. The duplicate and pixel-noise controls match the generated-reference input length, isolating the effect of generated content.

Input condition	Positive mIoU	Positive Acc@0.5
Original image only	42.68	48.67
Original + duplicate control	40.48	44.67
Original + pixel-noise control	44.42	49.73
Original + generated reference	33.44	35.33
Generated − duplicate	−7.04	−9.33
Generated − pixel noise	−10.98	−14.40

3.2 Naive diffusion-based augmentation harms visual grounding

In Fig. 1, we conduct experiments by inputting different samples into the grounding model. We then experiment with more controlled settings. Specifically, we compare four input conditions: the original image alone; a *duplicate control* that supplies the original image twice; a *pixel-noise control* that adds an equally sized but uninformative second image; and a *generated-reference* condition that pairs the original image with a diffusion-based image generated from the query. We use the frozen Qwen3.5-9B base model with greedy decoding and the fixed normalized-999 coordinate interpretation in all conditions. We sample data evenly across the four top-level categories in the GroundingME dataset, and report the performance in Tab. 1.

From Tab. 1, we can see that the generated reference shows the lowest performance. As the duplicate and pixel-noise controls have the same length of input image tokens as the generated reference, the performance degradation is therefore caused by the generated image itself: although the generated image is semantically related to the query, it does not suit the visual grounding task. Therefore, in the following subsections, we will study what information is really needed for this task.

3.3 Misaligned diffusion model can hardly help visual grounding

As visual grounding needs to spatially localize the target object, we first diagnose the diffusion-augmented model from the spatial evidence in its features. Let $H \in \mathbb{R}^{C\times h\times w}$ denote the feature of hidden states in a diffusion model. We use root mean square (RMS)

$$\mathrm{RMS}(H) = \sqrt{\frac{1}{Chw}\sum_{c=1}^{C}\sum_{i=1}^{h}\sum_{j=1}^{w} H_{cij}^2} \tag{3}$$

to control feature or latent magnitude. To construct the ***RMS-matched pure-noise*** control, we first apply the same scheduler and timestep to a zero image signal and the same sampled noise, obtaining a pure noise latent $\widetilde{z}_t$. We then rescale it to match the anchored latent z_t from Section 3.1:

$$z_t^{\mathrm{noise}} = \widetilde{z}_t \frac{\mathrm{RMS}(z_t)}{\mathrm{RMS}(\widetilde{z}_t) + \varepsilon}, \tag{4}$$

where ε is a small numerical constant. This control removes the source-image structure while holding the timestep, noise realization, and overall latent magnitude approximately fixed.

For spatial diagnosis, we convert H into a normalized response map by taking the channel-wise RMS at each position and normalizing over the map:

$$A_{ij}(H) = \frac{\sqrt{C^{-1}\sum_{c=1}^{C} H_{cij}^2}}{\sum_{u=1}^{h}\sum_{v=1}^{w}\sqrt{C^{-1}\sum_{c=1}^{C} H_{cuv}^2}}. \tag{5}$$

Let Ω_{GT} be the ground-truth box projected onto the $h\times w$ response grid, and let $a_{\mathrm{GT}} = |\Omega_{\mathrm{GT}}|/(hw)$ be its area fraction. The ***GT mass*** and ***GT lift*** are

$$M_{\mathrm{GT}}(H) = \sum_{(i,j)\in\Omega_{\mathrm{GT}}} A_{ij}(H), \qquad L_{\mathrm{GT}}(H) = \frac{M_{\mathrm{GT}}(H)}{a_{\mathrm{GT}}}. \tag{6}$$

A uniform response has $L_{\mathrm{GT}} = 1$, whereas a value above one means that the response is concentrated inside the target more strongly than expected from its area alone, indicating a better perception of the object for visual grounding. We report the results in Fig. 2.

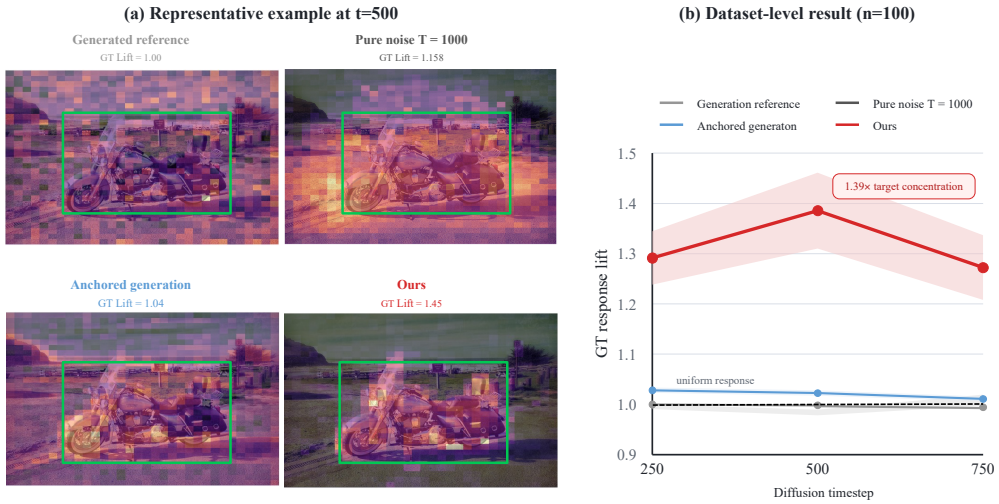


Figure 2: Green boxes denote ground-truth regions, and shaded bands show 95% confidence intervals. The naively generated reference produces a diffuse spatial response, whereas the pure-noise input happens to align with the model's target-centered focus. Anchoring the diffusion process to the input image further concentrates the response in the target region, highlighting the importance of alignment between diffusion and grounding.

From Fig. 2(left), we can see that the generated reference shows GT lift close to 1.0, meaning that these samples can hardly help the grounding model to spatially focus on the target object. Since the diffusion model and the visual grounding model are not initially designed to help each other, this result means that without explicitly aligning these two models, the grounding model can hardly find useful information from the diffusion model. In contrast, the pixel-noise control (the pure noise one) shows a higher GT lift value and higher performance in Tab. 1. As the noise tends to be applied in the center area of the image (see Appendix), this coincidentally aligns with the grounding model's focus on the target objects. Therefore, this inspires us to explicitly constrain the initial noise to be anchored on the query image. Specifically, in Eq. 2, we use the query q as the conditioned input as

$$F = D_{\phi}^{(\ell)}(z_t, t, q) \tag{7}$$

and use F for the subsequent generation. We term this kind of generation as *Image anchor + Cond* in Fig. 2, and we see that this simple operation increases the GT lift value.

In summary, to help the grounding model better perceive the input image, the diffusion model needs to be more explicitly aligned with the grounding model.

4 Method

4.1 Overview

Following the above analysis, we explicitly align the diffusion model with the grounding model. Specifically, we achieve this goal from both the side of the diffusion model and the side of the grounding model. For the diffusion side, we design an image-anchored teacher that extracts conditional and residual regional relations from a noise-corrupted latent of the input image; crop-and-pool Gram alignment transfers these relations to the student. For the grounding side, we design a contextual query-gated VLM student and use feature-space and bbox–null output-pair supervision on matched positive and hard-negative queries to calibrate target existence. At inference, only the trained VLM remains and predicts a box or `null`.

4.2 Contextual query-gated student

From a selected VLM layer, we extract visual tokens $V = \{V_i\}_{i=1}^{N}$ and the hidden state h_{ctx} at the last valid prompt position. The latter is a causal summary of the preceding image and query. A parameter-free, mean-preserving token gate computes

$$g_i = N \operatorname{softmax}_i\left(\frac{\cos(V_i, h_{\text{ctx}})}{\tau}\right), \qquad \widetilde{V}_i = [1 + \gamma(g_i - 1)]\, V_i, \tag{8}$$

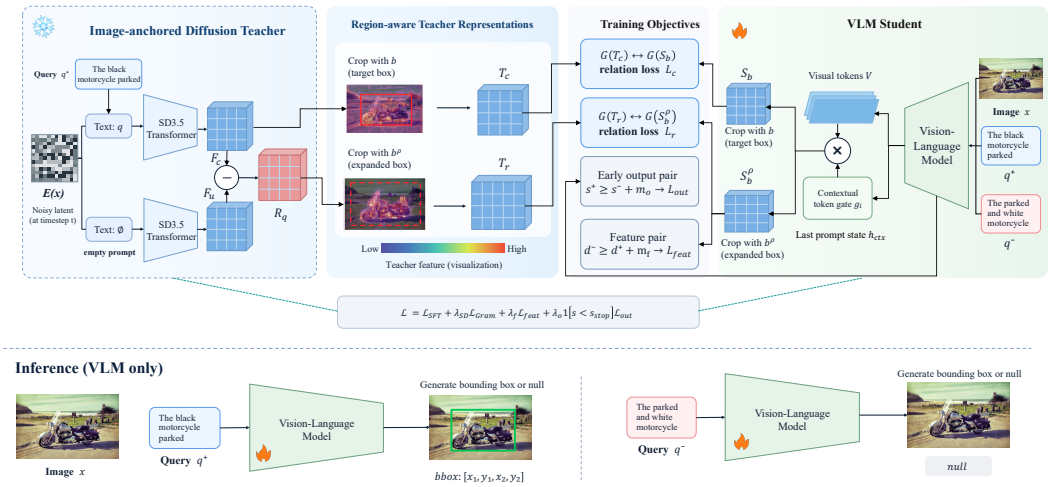


Figure 3: **Overview of DIRECT.** The image-anchored diffusion teacher provides conditional and residual regional relations. The VLM forms a spatial map by gating its visual tokens with the contextual state of the image–query prompt. Relational alignment supervises localization, while matched positive–negative queries impose feature-space and bbox–null margins. The teacher and all auxiliary objectives are removed at inference.

where τ controls selectivity and γ controls residual gate strength. We reshape $\widetilde{V}$ into a query-gated spatial map M_q.

4.3 REGION-AWARE RELATIONAL LOCALIZATION

We denote the "Crop with b" operation shown in Figure 3, followed by adaptive pooling, as the crop-and-pool operator $\mathcal{P}_b$. Let b^ρ expand b about its center by ratio ρ. The conditional branch uses the exact target box, whereas the residual branch uses its expanded context:

$$\begin{aligned} T_c &= \mathcal{P}_b(F_c), & T_r &= \mathcal{P}_{b^\rho}(R_q), \\ S_b(q) &= \mathcal{P}_b(M_q), & S_{b^\rho}(q) &= \mathcal{P}_{b^\rho}(M_q). \end{aligned} \tag{9}$$

For a pooled map $A \in \mathbb{R}^{P\times d}$, let $\widehat{A}$ contain its channel-centered, row-normalized spatial vectors. We compare spatial relations rather than channel values:

$$\begin{aligned} G(A) &= \widehat{A}\widehat{A}^\top, & D_G(A,B) &= \frac{\|G(A)-G(B)\|_F^2}{P^2}, \\ \mathcal{L}_{\text{Gram}} &= \frac{\mathcal{L}_c + \alpha\mathcal{L}_r}{1+\alpha}. \end{aligned} \tag{10}$$

Here $\mathcal{L}_c = D_G(S_b(q^+), T_c)$ and $\mathcal{L}_r = D_G(S_{b^\rho}(q^+), T_r)$ are the conditional and residual relation losses shown in Figure 3. This target is independent of teacher/student channel counts and transfers localization structure only for valid queries.

4.4 PAIRED EXISTENCE CALIBRATION

For the two relation branches, define $d_c^\pm = D_G(S_b(q^\pm), T_c)$ and $d_r^\pm = D_G(S_{b^\rho}(q^\pm), T_r)$. The feature-pair loss requires the counterfactual query to remain at least margin m_f farther from the same positive teacher relations:

$$\mathcal{L}_{\text{feat}} = \frac{[m_f + \text{sg}(d_c^+) - d_c^-]_+ + \alpha[m_f + \text{sg}(d_r^+) - d_r^-]_+}{1+\alpha}, \tag{11}$$

where $[a]_+ = \max(0,a)$ and sg is the identity in the forward pass but blocks gradients in the backward pass. We detach d^+ so the matched query remains a stable reference and the pair loss acts on the counterfactual branch; the positive branch is still trained by $\mathcal{L}_{\text{SFT}}$ and $\mathcal{L}_{\text{Gram}}$. This term is not used as an inference-time rejection score.

We also calibrate the first answer position that distinguishes a box from `null`. Let v_{bbox} and v_{null} denote the two divergent answer tokens and let ℓ_θ denote their logits. For a matched pair,

$$\begin{aligned} s^{\pm} &= \ell_\theta(v_{\text{bbox}} \mid x, q^{\pm}) - \ell_\theta(v_{\text{null}} \mid x, q^{\pm}), \\ \mathcal{L}_{\text{out}} &= \text{softplus}\big(m_o + s^{-} - \text{sg}(s^{+})\big)\,. \end{aligned} \tag{12}$$

The output pair uses the same asymmetry: s^+ is a detached reference, so the pair loss must reduce the counterfactual score s^- toward the null side, rather than satisfy the margin by further increasing the positive query's bbox preference. Thus m_f and m_o are training margins, not inference thresholds. The inequalities in Figure 3 show the desired margins, while the expressions above are their training surrogates.

4.5 Joint objective and inference

At optimizer update u, the full objective is

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda_{\text{SD}}\mathcal{L}_{\text{Gram}} + \lambda_f\mathcal{L}_{\text{feat}} + \lambda_o \mathbb{1}[u < u_{\text{stop}}]\mathcal{L}_{\text{out}}. \tag{13}$$

The first two terms form DiRect-Loc; DiRect additionally includes feature- and output-level pairing. Only $\mathcal{L}_{\text{out}}$ is disabled after u_{stop}, while $\mathcal{L}_{\text{feat}}$ remains active. The diffusion teacher is frozen and cached offline. At inference, the teacher, ground-truth regions, and auxiliary objectives are removed; the trained VLM directly generates a bounding box or `null`.

5 Experiments

5.1 Setup

Training uses 5K examples: 3,500 FineCops positives, 750 FineCops hard negatives (Liu et al., 2024), and 750 gRefCOCO no-target samples (He et al., 2023). We evaluate transfer on the held-out FineCops test split and gRefCOCO no-target val/test splits; GroundingME (Li et al., 2026) serves as the out-of-domain benchmark. We report mIoU and Acc@0.5 for localization, rejection accuracy on no-target queries, and protocol-specific balanced metrics. Unless stated otherwise, experiments use Qwen3.5-9B trained for one epoch. Teacher features are cached offline, and confidence intervals are computed by paired bootstrap over query pairs.

5.2 Main GroundingME results

Table 2: Main GroundingME comparison. † marks published GroundingME baselines (Li et al., 2026); remaining rows are our Qwen3.5-9B runs. Baseline mIoU/mAcc are unreported (shown as –). Our results use `official-auto`. Bold marks the best fine-tuned row; Δ is relative to SFT.

Model / objective	mIoU	mAcc	Acc@0.5				
			Overall	Discr.	Spatial	Limited	Reject
InternVL3.5-8B†	–	–	3.3	6.4	4.0	1.7	1.5
Llama-4-Maverick†	–	–	13.0	18.1	22.3	5.0	6.0
Qwen2.5-VL-72B†	–	–	29.6	48.5	40.3	23.7	3.0
Qwen3-VL-4B†	–	–	33.9	56.4	28.3	47.0	0.0
Qwen3-VL-8B†	–	–	31.0	61.3	26.3	36.0	0.0
Qwen3-VL-A3B†	–	–	35.7	63.2	30.0	46.7	0.0
GLM-4.5V†	–	–	32.1	52.9	42.0	29.3	0.5
Qwen3-VL-32B†	–	–	39.5	75.0	47.3	34.0	0.0
Gemini-2.5-Pro†	–	–	20.7	34.8	34.0	7.0	7.0
Seed-1.6-Vision†	–	–	42.6	59.8	58.7	42.7	1.0
Qwen3-VL-A22B†	–	–	45.1	69.6	49.7	54.0	0.0
Qwen3.5-9B	34.79	30.13	37.61	60.29	33.00	52.00	0.00
SFT	39.51	36.79	41.89	64.22	37.33	36.00	34.83
DiRect-Loc (SD-GRAM)	**41.84**	**38.89**	**44.58**	**69.61**	**39.67**	**41.67**	30.85
DiRect	41.43	**38.89**	44.18	65.69	38.33	40.67	**36.32**
Δ *(vs. SFT)*	*+1.92*	*+2.10*	*+2.29*	*+1.47*	*+1.00*	*+4.67*	*+1.49*

Table 2 places our controlled Qwen3.5-9B comparison alongside representative results reported by GroundingME. Within the controlled comparison, DIRECT-LOC improves SFT by 2.33 mIoU and 2.69 Acc@0.5. DIRECT combines DIRECT-LOC with feature- and output-level pairing. It improves SFT by 1.92 mIoU and 2.29 Acc@0.5 while increasing rejection accuracy from 34.83 to 36.32. Importantly, it also improves every positive GroundingME category over SFT: Discriminative, Spatial, and Limited Acc@0.5 increase by 1.47, 1.00, and 4.67 points, respectively. This makes DIRECT the balanced operating point, whereas DIRECT-LOC remains the localization-oriented operating point. Appendix A.1 reports the complete level-2 breakdown.

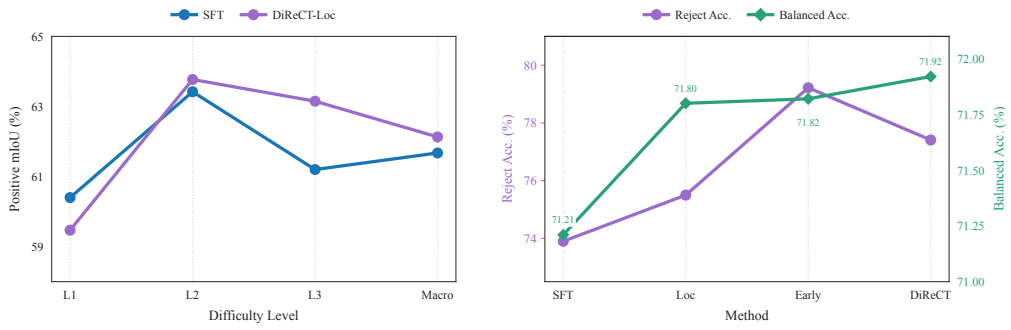
5.3 TRANSFER LOCALIZATION AND REJECTION

Table 3: FineCops held-out transfer under the forced-box positive-localization diagnostic. Every target-present query is required to produce a box, isolating localization quality from target-existence decisions. Bold denotes the best result in each metric, and blue shading identifies the full DIRECT model.

Method	Positive mIoU ↑				Acc.@0.5 ↑
	All	L1	L2	L3	
Base	70.98	75.75	64.52	59.56	79.95
SFT	74.06	78.12	68.41	65.53	82.35
DIRECT-LOC (SD-GRAM)	**74.71**	78.74	**68.95**	**67.34**	83.07
DIRECT-LOC + Early Output Pair	74.65	**78.90**	68.69	66.04	**83.15**
DIRECT	74.12	78.39	68.15	65.32	82.60

Figure 4 summarizes standard bbox-or-null transfer. Compared with SFT, DIRECT-LOC trades a 0.93-point decrease on L1 for gains on L2 (0.35) and L3 (1.95), yielding a 0.46-point macro improvement. This pattern indicates that its standard-protocol gains are concentrated on harder queries. Early Output Pair attains the highest negative rejection accuracy (79.22), whereas DIRECT achieves the highest balanced accuracy (71.92) while retaining 77.41 rejection accuracy.

Table 3 requires a box for every positive query, isolating localization. DIRECT-LOC raises forced-box mIoU from 74.06 (SFT) to 74.71, with its largest gain on L3 (65.53 to 67.34); DIRECT-LOC + Early Output Pair reaches 74.65 and full DIRECT 74.12. Rejection and balanced accuracy are assessed separately under the standard bbox-or-null protocol (Figure 4), where DIRECT achieves the highest balanced accuracy.



(a) Localization by difficulty

(b) Rejection and balanced accuracy

Figure 4: **Standard FineCops transfer under the bbox-or-null protocol.** (a) Positive mIoU for SFT and DIRECT-LOC across difficulty levels and their macro average. (b) Rejection and balanced accuracy for SFT, DIRECT-LOC, DIRECT-LOC + Early Output Pair, and DIRECT.

5.4 Ablations and controlled tests

Table 4: Component ablation on GroundingME. Pos. mIoU is computed only over target-present queries and is reported alongside rejection accuracy to separate localization from target-existence decisions.

Variant	mIoU	Pos. mIoU	Acc@0.5	mAcc	Reject Acc.
Anchor only	39.95	44.46	42.79	37.17	21.89
Residual only	37.87	41.62	40.60	35.00	22.89
DiReCT w/o Query Gate	40.44	42.22	42.99	37.95	33.33
Output Pair only	39.94	42.96	42.89	37.09	27.86
Shuffled teacher	39.59	41.02	42.09	37.05	**33.83**
DiReCT-Loc (SD-GRAM)	**41.84**	**44.59**	**44.58**	**38.89**	30.85

Table 5: Ablation of the two paired objectives on Qwen3.5-9B. Category columns report mIoU, except Reject, which reports rejection accuracy. A is DiReCT-Loc (SD-GRAM); B adds feature pairing; C adds early output pairing; and D is DiReCT. DiReCT improves SFT on the aggregate score and all four categories.

ID	Objective	Overall mIoU	Acc@0.5	Discr.	Spatial	Limited	Reject
–	SFT	39.51	41.89	59.50	36.23	32.33	34.83
A	DiReCT-Loc (SD-GRAM)	**41.84**	**44.58**	**64.09**	**38.57**	**37.34**	30.85
B	DiReCT-Loc + feature pair	40.38	43.18	61.73	37.07	35.58	30.85
C	DiReCT-Loc + output pair	41.05	43.58	60.14	38.46	33.82	**36.32**
D	DiReCT	41.43	44.18	60.81	36.74	36.36	**36.32**

Table 5 shows complementary effects. Feature pairing alone (B) leaves rejection unchanged at 30.85 and lowers overall mIoU from 41.84 to 40.38. Adding the output pair (C) raises rejection to 36.32, a 5.47-point gain over A, while reducing overall mIoU by 0.79 (41.84 to 41.05). Combining both (D) preserves C's rejection accuracy, recovers 0.38 overall mIoU, and improves Limited mIoU from 33.82 to 36.36. Together, these results indicate that output pairing increases rejection accuracy but introduces a localization trade-off; adding Feature Pair in the combined objective preserves the rejection gain while recovering some localization, most clearly on Limited queries.

Additional analyses. Forced-box mIoU transfers from SFT to DiReCT on LISA (Lai et al., 2024) (65.09 to 65.79) and RefCOCOg-test (Mao et al., 2016) (82.50 to 82.56); standard bbox-or-null scores also reflect abstention on these target-present sets. The appendix reports trajectory, SD-GRAM sensitivity, and teacher-feature diagnostics.

6 Limitations and Broader Impact

Offline diffusion-feature extraction adds data-preparation and storage cost, but the teacher is not used at inference. We evaluate Qwen3.5-9B with SD3.5; extending to other teachers and grounding backbones remains future work.

7 Conclusion

Pure generation can preserve query semantics yet disrupt the scene geometry needed for grounding. Image-anchored diffusion residuals provide spatial evidence, while paired diagnostics show that localization and query validity require distinct objectives. DiReCT combines relational diffusion supervision, feature separation, and early bbox–null calibration. On GroundingME, DiReCT-Loc improves mIoU over SFT, while full DiReCT improves rejection accuracy and retains stronger localization; forced-box transfer also improves on FineCops, LISA, and RefCOCOg. The teacher thus supplies localization evidence during training; inference uses only the VLM.

AI USE STATEMENT

Generative AI tools were used only to aid and polish the writing of this paper, including improvements to clarity, grammar, and presentation. We reviewed and revised all AI-assisted text and take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

Sections 3–4 describe the problem formulation, method, and training objectives, while Section 5 documents the datasets, evaluation protocols, and principal experimental settings. The appendix provides the explored trajectory variants, SD-GRAM weight sensitivity, and teacher-feature diagnostics. The supplementary material includes data-preparation utilities, training and evaluation scripts, configuration files, and commands for reproducing the reported tables and figures. We will publicly release the code and configuration files upon publication.

REFERENCES

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *Proceedings of ICCV*, pp. 1769–1779, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of ICML*, 2024.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized referring expression comprehension. In *Proceedings of ICCV*, pp. 23336–23345, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of ICCV*, pp. 1780–1790, 2021.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of EMNLP*, pp. 787–798, 2014.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of CVPR*, pp. 9579–9589, 2024.

Rang Li, Lei Li, Shuhuai Ren, et al. Groundingme: Exposing the visual grounding gap in MLLMs through multi-dimensional evaluation. In *Proceedings of CVPR*, 2026.

Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. In *Proceedings of EMNLP*, pp. 15440–15457, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Run Luo, Yunshui Li, Longze Chen, et al. DEEM: Diffusion models serve as the eyes of large language models for image perception. In *Proceedings of ICLR*, 2025.

Shijie Ma, Yuying Ge, Teng Wang, Yuxin Guo, Yixiao Ge, and Ying Shan. Genhancer: Imperfect generative models are secretly strong vision-centric enhancers. *arXiv preprint arXiv:2503.19480*, 2025.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of CVPR*, pp. 11–20, 2016.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of CVPR*, pp. 3967–3976, 2019.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *Proceedings of ICLR*, 2024.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, et al. GLaMM: Pixel grounding large multimodal model. In *Proceedings of CVPR*, pp. 13009–13018, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR*, pp. 10684–10695, 2022.

Adriana Romero, Nicolas Ballas, Samira Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *Proceedings of ICLR*, 2015.

Zhen Sun, Yunhang Shen, Jie Li, Xing Sun, Pingyang Dai, Liujuan Cao, and Rongrong Ji. DS-VLM: Diffusion supervision vision language model. In *Proceedings of ICML*, pp. 57667–57679, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *Proceedings of ICLR*, 2020.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of ICCV*, pp. 1365–1374, 2019.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. ODISE: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of CVPR*, pp. 15695–15706, 2023.

Haoxuan You, Haotian Zhang, Zhe Gan, et al. Ferret: Refer and ground anything anywhere at any granularity. In *Proceedings of ICLR*, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander Berg, and Tamara Berg. Modeling context in referring expressions. In *Proceedings of ECCV*, pp. 69–85, 2016.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of ICLR*, 2017.

Yuliang Zhan, Xinyu Tang, Han Wan, Jian Li, Ji-Rong Wen, and Hao Sun. L2V-CoT: Cross-modal transfer of chain-of-thought reasoning via latent intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Tiezheng Zhang, Yitong Li, Yu-Cheng Chou, Jieneng Chen, Alan Yuille, Chen Wei, and Junfei Xiao. Vision-language-vision auto-encoder: Scalable knowledge distillation from diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of ICCV*, pp. 5729–5739, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Additional Analyses

A.1 Fine-grained GroundingME breakdown

Table 6 resolves the three positive-query groups in Table 2 into their eight level-2 subtasks and separately reports the four rejection subtypes. This view distinguishes the localization-oriented behavior of DiReCT-Loc from the more balanced localization–rejection behavior of DiReCT.

Table 6: Fine-grained GroundingME breakdown for Qwen3.5-9B. The upper panel reports positive-query Acc@0.5 and the lower panel reports rejection accuracy. Discriminative comprises Appearance, Component, State, and Text; Spatial comprises Counting and Relationship; Limited comprises Occlusion and Small. Bold marks the best result among all four rows in each column.

Method	Appearance	Component	State	Text	Counting	Relationship	Occlusion	Small
Base	57.69	72.00	42.31	70.00	24.67	41.33	63.33	**40.67**
SFT	63.46	74.00	50.00	70.00	28.00	46.67	64.67	7.33
DiReCT-Loc	**65.38**	**84.00**	**55.77**	**74.00**	28.67	**50.67**	69.33	14.00
DiReCT	63.46	74.00	51.92	**74.00**	**29.33**	47.33	**70.00**	11.33

Method	Appearance	Component	State	Text	Overall
Base	0.00	0.00	0.00	0.00	0.00
SFT	34.00	35.29	24.00	**46.00**	34.83
DiReCT-Loc	32.00	33.33	24.00	34.00	30.85
DiReCT	**38.00**	**37.25**	**28.00**	42.00	**36.32**

DiReCT-Loc provides the strongest localization result on five of the eight positive subtypes and improves over SFT on every subtype. Adding paired calibration preserves or improves SFT on all eight positive subtypes while raising overall rejection accuracy from 34.83 to 36.32, with gains on Appearance, Component, and State rejection. The remaining gap on Small objects relative to Base is shared by the fine-tuned models, identifying small-object retention as a limitation of the current training mixture rather than a consequence specific to output pairing.

A.2 Exploratory diffusion-trajectory supervision

Before adopting the fixed-timestep relational target used by SD-GRAM, we explored whether a VLM could instead learn the denoising trajectory itself. These variants attach a spatial prediction head to the VLM and supervise the direction of the diffusion feature displacement at one or more timesteps. We tested (i) a single-timestep target at $t = 550$, (ii) a three-timestep target at $t \in \{850, 550, 250\}$ that was applied only during the first 32 updates and then followed by ordinary SFT, and (iii) a five-timestep target at $t \in \{950, 750, 550, 350, 150\}$ with a smaller trajectory weight. Table 7 reports every saved checkpoint from these runs.

Table 7: Comparison of exploratory diffusion-trajectory supervision with the fixed-timestep SD-GRAM reference on GroundingME using the `official-auto` coordinate interpretation.

Variant	Timesteps / schedule	λ_{traj}	Ckpt.	Coord.	mIoU	Acc@0.5	Discr.	Spatial	Limited	Reject
1T trajectory	550, full run	0.05	53	auto	37.88	40.90	67.16	38.00	48.00	7.96
1T trajectory	550, full run	0.05	105	auto	**39.66**	**42.09**	62.75	35.00	41.00	33.33
3T trajectory → SFT	850/550/250, stop 32	0.05	32	auto	38.83	41.49	**68.63**	**40.67**	40.33	16.92
3T trajectory → SFT	850/550/250, stop 32	0.05	63	auto	39.20	41.59	67.16	38.67	36.33	27.86
5T trajectory	950/750/550/350/150, full run	0.02	32	auto	38.65	40.70	55.39	33.33	37.00	**42.29**
5T trajectory	950/750/550/350/150, full run	0.02	63	auto	39.36	41.69	65.20	37.00	38.00	30.35
Fixed SD-GRAM (main)	$t = 500$, relational Gram	–	63	auto	41.84	44.58	69.61	39.67	41.67	30.85

The trajectory objective produced unstable localization–rejection trade-offs rather than a uniform improvement. For example, the early 5T checkpoint reached 42.29% rejection accuracy, but its Discriminative accuracy fell to 55.39%. The 3T schedule briefly improved Spatial accuracy to 40.67%, yet the gain did not persist after the trajectory loss was disabled. Across the trajectory runs, the best mIoU and Acc@0.5 were 39.66 and 42.09, respectively, both below the final fixed-timestep SD-GRAM result (41.84 and 44.58). Moreover, the direction-prediction loss remained approximately

0.93–0.96 near the end of the full trajectory runs, consistent with a noisy, high-dimensional target that is difficult to fit in the short 5K-data regime.

These observations motivated a simpler design: cache image-anchored diffusion features at a fixed timestep and transfer their regional relational structure, rather than asking the VLM to reproduce the full denoising dynamics. Across the explored settings, trajectory supervision did not outperform the simpler fixed-feature SD-GRAM objective. We therefore adopt SD-GRAM in the main method and report the trajectory variants as a design exploration.

A.3 Sensitivity to the SD-GRAM weight

We next vary only the outer SD-GRAM loss weight λ_{SD}. All other factors are fixed: SD3.5 Medium features from `transformer_blocks.12` at $t = 500$, a 16×16 Gram map, a 1.5-times residual-region expansion, contrast coefficient 0.5, and the same query gate, optimizer, data, and seed. Table 8 reports the archived GroundingME evaluations for the retained weight settings.

Table 8: Sensitivity to the SD-GRAM loss weight on GroundingME. Positive accuracy excludes rejection cases. The results show a non-monotonic response, with moderate weighting providing a strong balance between localization and overall accuracy.

	Checkpoint 32			Checkpoint 63			
λ_{SD}	mIoU	Overall Acc.	Reject Acc.	mIoU	Overall Acc.	Pos. Acc.	Reject Acc.
0.01	40.84	43.08	47.76	**41.87**	44.18	45.15	**40.30**
0.03	40.75	42.49	**70.15**	41.60	44.18	46.39	35.32
0.10	37.65	39.50	54.23	40.19	42.79	45.90	30.35
0.30	**42.32**	**44.38**	64.68	40.21	42.79	46.52	27.86
0.50	40.64	43.48	25.37	41.84	**44.58**	**48.01**	30.85
0.80	38.30	40.00	59.70	40.54	42.99	44.53	36.82
1.50	38.42	40.40	55.22	40.59	43.18	46.27	30.85
2.00	38.54	40.10	63.68	39.02	41.29	43.41	32.84

The sweep reveals two effects. First, the response is non-monotonic rather than "larger is better." At checkpoint 63, increasing the weight beyond 0.5 reduces positive accuracy from 48.01% to 44.53% at 0.8 and 43.41% at 2.0, consistent with over-regularizing the answer-token SFT objective. Second, aggregate scores at an intermediate checkpoint can hide a different operating point. For example, the 0.3/checkpoint-32 model obtains 44.38% overall accuracy largely through 64.68% rejection accuracy, while its positive accuracy is only 39.30%. In contrast, the 0.5 configuration maintains 48.01% positive accuracy at both checkpoints and improves rejection accuracy from 25.37% to 30.85%. We therefore use 0.5 as the representative SD-GRAM weight: it provides the strongest final positive localization and overall accuracy without relying on an unusually rejection-heavy intermediate checkpoint.

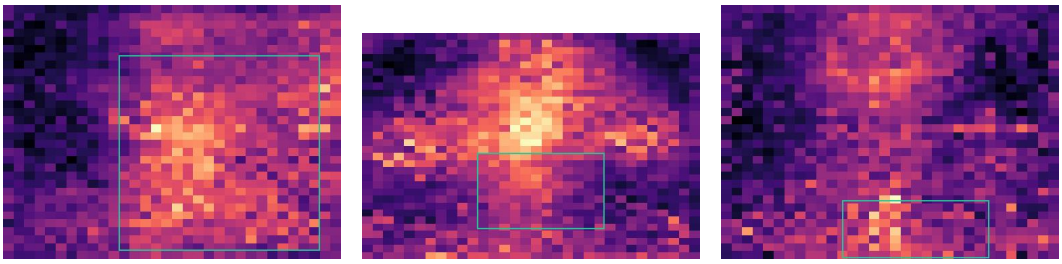
A.4 Teacher-feature diagnostic statistics

Table 9: Teacher-feature diagnosis on paired FineCops examples. The feature-change ratio measures sensitivity to a query swap, while the GT-lift gap measures whether that change preferentially suppresses the hard-negative response.

Latent	Feature	Correct GT lift	Change ratio	GT-lift gap
Pure noise	Cond.	0.998	0.021	−0.000
Pure noise	Residual	0.984	0.560	−0.003
Image anchor	Cond.	1.022	0.035	0.000
Image anchor	Residual	**1.386**	**0.547**	0.011

A.5 Pure-noise endpoint response examples

Figure 5 shows query-conditioned SD3.5 response maps at the terminal pure-noise timestep ($t = T = 1000$) for three selected FineCops negative examples. The maps remain spatially nonuniform at this endpoint; the green outlines mark the annotated regions used to calculate GT lift. These examples are qualitative and are not an aggregate estimate of localization performance.



Case 004 GT lift: 1.102
The elephant, that is large, to the right of the tan rope.

Case 007 GT lift: 1.043
The keyboard that has the same color with the bar stool.

Case 009 GT lift: 1.129
The large, parked and blue car to the right of the tree trunk and to the left of the directional map.

Figure 5: Query-conditioned response maps at the pure-noise endpoint for three FineCops negative examples. Green outlines indicate the annotated regions. The GT-lift values are case-specific and are reported for these examples only.