COUNTERGROUND: DIAGNOSING AND REPAIRING COUNTERFACTUAL FAILURES IN VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual grounding is typically evaluated by box overlap, which conflates association with the intended instance and localization precision. We introduce **CounterGround**, a paired counterfactual benchmark that fixes the image, object category, and candidate set while a spatial relation edit switches the annotated target instance. Across 5,635 pairs from remote-sensing and natural-image domains, recent multimodal large language models (MLLMs) exhibit a consistent behavioral separation: predictions can respond consistently to the relation edit and remain associated with the intended instances while failing the paired intersection-over-union (IoU) criterion. This separation persists as Qwen3.5 scales from 2B to 9B, despite higher grounding accuracy. Controlled Diagnosis on Qwen3.5-2B and Qwen3-VL-2B-Instruct shows that selecting among explicit ground-truth candidates is substantially easier than free-form box localization, while fixed-center extent prediction and a ground-truth-size oracle leave substantial residual error. Guided by this diagnosis, joint geometric refinement raises Pair Accuracy on CounterGround-DIOR validation from 48.65% to 56.06%, and adding local visual-language features raises it to 61.33%. The repair formulation also improves the reported point estimates on held-out CounterGround-DIOR with both backbones and on CounterGround-COCO, RefCOCO, and VRSBench under their specified training protocols. CounterGround links controlled target-switch evaluation to behavioral diagnosis and targeted repair.

1 INTRODUCTION

Visual grounding is usually summarized by whether a predicted box sufficiently overlaps the referred object. This outcome can conflate distinct behaviors: a model may respond consistently to a relation edit that changes the target and associate its prediction with the intended instance, yet produce an inaccurate box. As visual grounding becomes a capability of general-purpose multimodal large language models (MLLMs), we ask: *when a controlled relation edit switches the target, which observable behaviors remain correct even when precise box localization fails?*

Standard grounding benchmarks score image–query examples independently, without explicitly evaluating controlled relationships between queries about the same scene. Replacing `leftmost` with `rightmost` in a referring expression switches the annotated target to a different instance of the same category. Independent scores do not characterize how relation-consistent responses, intended-instance association, and precise localization relate within this controlled target switch (Figure 1).

We study this problem through **CounterGround**, a paired counterfactual framework in which two expressions share the same image, semantic category, and candidate set, while a controlled relation edit changes the annotated target to a different physical instance. The primary evaluation comprises 4,613 held-out CounterGround-DIOR pairs and 1,022 CounterGround-COCO pairs, totaling 5,635 pairs (11,270 samples) across remote sensing and natural image domains. A deterministic audit verifies the predefined construction invariants for all primary evaluation pairs. CounterGround focuses on predefined spatial target switches; other forms of counterfactual referring-expression reasoning fall outside its scope.

Across models, CounterGround reveals a consistent separation: predicted centers can move in the correct direction and remain associated with the intended same-class instances even when the boxes fail the paired IoU criterion. Scaling Qwen3.5 from 2B to 9B improves absolute localization accuracy

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.

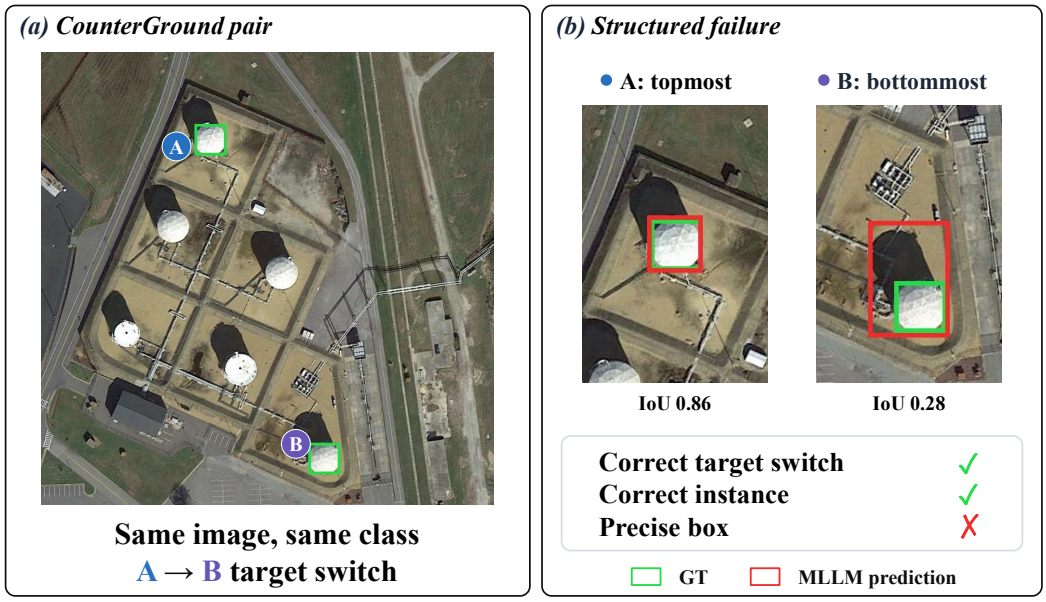

Figure 1: **CounterGround pair and structured behavioral separation.** A controlled target-defining relation edit switches the annotated physical referent while preserving the scene and object class. Here, the model responds consistently to the relation change and remains associated with the intended instances, yet only one member satisfies the IoU criterion. Green boxes denote ground truth and red boxes denote MLLM predictions.

in both domains but does not remove this separation. The central finding concerns the behaviors that remain correct when localization fails, beyond the expected gap between Sample and Pair Accuracy.

Controlled Diagnosis probes the remaining localization difficulty. With explicit ground-truth (GT) candidates, target-instance selection achieves substantially higher Pair Accuracy than free-form grounding (88.44% vs. 48.65% for Qwen3.5-2B). Independently estimated centers are frequently associated with the intended instance, yet fixed-center extent prediction and a GT-size oracle leave substantial residual error. These interventions reveal a gap between assisted target selection and precise localization.

This diagnosis motivates a focused repair. Joint center-extent refinement raises Pair Accuracy on CounterGround-DIOR validation from 48.65% to 56.06%, and adding local visual-language features raises it to 61.33%; a matched direct-crop control performs substantially worse. The formulation also improves the reported point estimates on held-out CounterGround-DIOR with both backbones, on CounterGround-COCO, and under the native RefCOCO and VRSBench evaluation protocols, using the specified training protocols for each dataset.

Our contributions are threefold. First, **CounterGround** reveals differences between relation-consistent responses, intended-instance association, and precise localization through controlled paired target switches. Second, **Controlled Diagnosis** identifies a substantial gap between target selection with explicit GT candidates and free-form localization. Third, a **diagnosis-guided repair** combines joint geometric refinement with local visual-language features to improve paired and native grounding point estimates across held-out, cross-backbone, and cross-dataset evaluations.

2 Related Work

Recent multimodal language models increasingly integrate visual grounding as a general-purpose referential capability. Systems such as Kosmos-2, Ferret, LLaVA-Grounding, and GroundingGPT connect natural-language interaction to regions, points, masks, or bounding boxes (Peng et al., 2024; You et al., 2024; Zhang et al., 2024; Li et al., 2024b). Ren et al. (2026) study GETok, an alternative

tokenization of spatial outputs for autoregressive grounding. Localization-oriented architectures, including MDETR, GLIP, and Grounding DINO, remain strong specialist foundations (Kamath et al., 2021; Li et al., 2022; Liu et al., 2024b). More recent work examines whether semantic reasoning and metric spatial prediction should share the same mechanism. Wu et al. (2025) use F-LMM to provide frozen MLLM features to downstream spatial modules, whereas Kang et al. (2026) explicitly disentangle high-level reasoning from box prediction. These approaches primarily improve how referred content is represented or localized. CounterGround studies the reliability of these behaviors: it asks how relation-consistent responses, intended-instance association, and precise box localization diverge under a controlled target switch in a fixed image.

A complementary line of work makes grounding evaluation more diagnostic by testing compositional reasoning, distractor handling, robustness, and invalid premises. CLEVR-Ref+, Cops-Ref, and FineCops-Ref provide diagnostic settings for compositional referring-expression reasoning (Liu et al., 2019; Chen et al., 2020; Liu et al., 2024a). Dong et al. (2026) introduce Ref-Adv to reduce shortcut solutions with harder distractors and richer expressions, while Li et al. (2026) use GroundingME to expose MLLM grounding weaknesses through multidimensional evaluation. The most closely related settings involve counterfactual and false-premise referring tasks. Yu & Li (2024) construct C-RefCOCO/+/g by editing attributes and jointly predict localization and referent existence; Wu et al. (2024) study absent-target premises in See, Say, and Segment; and Lombardo et al. (2026) analyze grounding under controlled object- and context-level replacements that create image-text mismatches. CounterGround uses a *positive-positive target switch*: the paired expressions share the same image, semantic category, and candidate set, while a minimal target-defining relation edit changes the annotated target to a different physical instance.

Our diagnosis complements modular semantic-spatial designs: paired evaluation reveals differences among observable behaviors, and controlled conditions probe target selection and localization without assuming a particular internal MLLM computation. Local visual search and refinement provide precedents for repair. Wu & Xie (2024) use guided visual search in V*, Li et al. (2025) dynamically focus on local regions with DyFo without additional training, and Zhang et al. (2026) apply GeoViS to remote-sensing visual grounding. CounterGround compares joint geometric refinement, the addition of local visual-language features, and a matched direct-crop control to examine their contributions to repair.

3 COUNTERGROUND BENCHMARK

CounterGround organizes grounding examples into *counterfactual pairs*. The two members share an image, object category, and same-class candidate set, but a minimal relation edit changes the annotated target instance.

3.1 PAIRED TASK AND CONSTRUCTION

Let $x = (I, q, b)$ denote a grounding sample, where I is an image, q is a referring expression, and b is the target box. The evaluation unit is

$$p_i = \left[(I_i, q_i^{(a)}, b_i^{(a)}), (I_i, q_i^{(b)}, b_i^{(b)}) \right]. \tag{1}$$

The queries differ by a controlled relation transformation $q_i^{(b)} = \mathcal{T}(q_i^{(a)})$ that must change the target identity: $z_i^{(a)} \neq z_i^{(b)}$, where $z_i^{(a)}$ and $z_i^{(b)}$ identify the two physical instances.

The frozen benchmark uses two relation families:

$$\texttt{leftmost} \rightarrow \texttt{rightmost}, \quad \texttt{topmost} \rightarrow \texttt{bottommost}. \tag{2}$$

Each pair changes exactly one target-defining relation. The original relation token must occur exactly once, and only that token is replaced. Both annotated targets must be distinct instances in the shared same-class candidate set. This construction links a minimal linguistic edit to a required change in target identity.

We construct CounterGround from DIOR (Li et al., 2020) and COCO (Lin et al., 2014). The held-out CounterGround-DIOR test set contains 4,613 pairs (9,226 samples) from the DIOR test split used in

our protocol; CounterGround-COCO contains 1,022 pairs (2,044 samples) from the corresponding COCO test split. Together, the two test sets contain 5,635 pairs and 11,270 samples. A separate CounterGround-DIOR validation set contains 2,298 pairs (4,596 samples) from the DIOR validation split and supports Controlled Diagnosis and the primary repair comparison. Diagnosis and repair development are thus separate from held-out test evaluation.

3.2 Evaluation and Integrity

All CounterGround localization accuracies use an IoU threshold of $\tau = 0.5$; invalid or missing predictions receive an IoU of zero. *Sample Accuracy* scores individual image-query examples. *Pair Accuracy* counts a pair as correct only if both members satisfy $\mathrm{IoU} \geq 0.5$. Both metrics include all evaluated examples or pairs. Because Pair Accuracy is inherently stricter, its gap from Sample Accuracy alone does not establish counterfactual unreliability. We instead examine which behaviors remain correct when localization fails.

Three additional diagnostics are computed only over pairs with two valid predictions. *Direction Accuracy* measures whether the predicted center moves along the relevant axis in the same direction as the GT target switch. *Instance Association Accuracy* assigns each predicted center to its nearest annotated same-class candidate and requires both assignments to match the intended target identities. *ID✓/IoU×* is the fraction of valid pairs for which both instance assignments are correct but at least one box fails the IoU criterion. These diagnostics distinguish relation-consistent response, instance association, and box precision; their denominator differs from that of Pair Accuracy.

We run a deterministic construction audit over all 5,635 test pairs. It checks: (1) exactly two uniquely identified members; (2) the same source image and dimensions; (3) the same object category; (4) a predefined relation transition; (5) an edit restricted to the designated relation; (6) distinct target instances; (7) finite, ordered target, counterpart, and candidate boxes; and (8) a shared same-class candidate set containing both targets. Every pair passes all eight checks. The audit uses only frozen annotations and construction metadata, without inspecting model predictions or repair outcomes. It therefore verifies structural integrity independently of model performance.

4 Controlled Diagnosis and Diagnosis-Guided Repair

A failed box does not reveal whether the difficulty lies in target selection, center localization, extent estimation, or their interaction. We probe these factors under controlled prediction conditions and use the results to guide repair. We parameterize a box as $b = (\mathbf{c}, \mathbf{s})$, with center $\mathbf{c} = (c_x, c_y)$ and extent $\mathbf{s} = (w, h)$, equivalent to the corner coordinates used for the evaluation.

4.1 Controlled Diagnosis

We evaluate four conditions on the same CounterGround-DIOR validation set. In *free-form grounding*, the MLLM receives an image and query and directly predicts a box. In *GT-candidate selection*, we render all same-class GT candidate boxes with randomly permuted labels that do not encode spatial position, and ask the same MLLM to return the target label. We map the selected label back to its annotated candidate and average results over three deterministic label permutations. This condition measures target selection with supplied GT geometry and is therefore a diagnostic intervention.

The other two conditions use an independently estimated center. A Stage-1 model fitted only on the DIOR training split predicts a spatial candidate distribution, whose expectation gives a fixed center $\widetilde{\mathbf{c}}_i$. In *fixed-center extent prediction*, a separate predictor estimates width and height around that center; results are averaged over three predeclared training seeds. The *GT-size oracle* uses the same center with the target's GT width and height. These conditions probe extent prediction and the error remaining when extent is exact. They characterize observable behavior without decomposing the MLLM's internal computation. Figure 2 illustrates all four conditions on one validation pair.

4.2 Diagnosis-Guided Repair

We compare three repair conditions: *center-only correction*, *joint geometry refinement*, and *joint geometry + local evidence* (full repair). Center-only correction combines the learned center readout

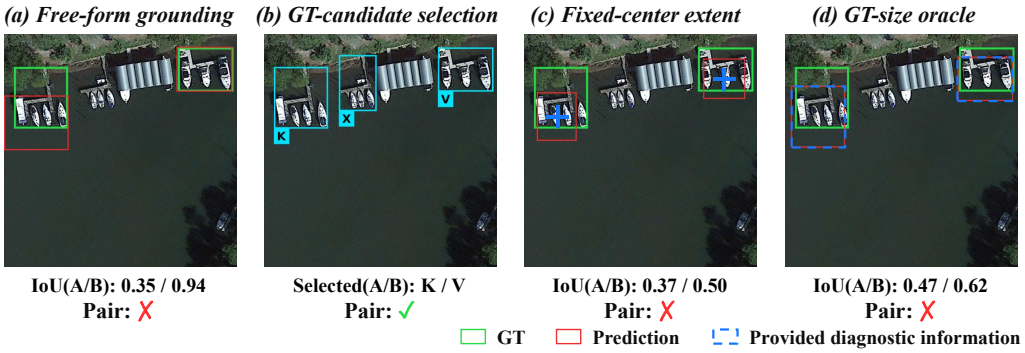

Figure 2: **Controlled Diagnosis on a shared validation pair.** (a) Free-form grounding directly predicts boxes. (b) GT-candidate selection explicitly supplies same-class ground-truth candidate boxes and asks which instance is referred to. (c) Fixed-center extent prediction uses the frozen Stage-1 center (blue +) and predicts only extent. (d) The GT-size oracle retains the same center while supplying ground-truth width and height (blue dashed boxes). Supplied GT geometry is diagnostic information only, not part of a deployable grounding system. Green boxes denote ground truth and red boxes denote predictions.

with the same extent-prediction family used in the repair pipeline, fitted only on training data. A matched direct-crop control tests whether using the crop prediction directly is sufficient.

The *joint geometry refinement* and *joint geometry + local evidence* conditions start from the frozen joint-geometry ensemble box $\bar{b} = (\bar{\mathbf{c}}, \bar{\mathbf{s}})$. A square crop is centered on $\bar{\mathbf{c}}$ with nominal side length

$$\rho = \mathrm{clip}\big(4\max(\bar{w}, \bar{h}),\ 0.20,\ 0.60\big)\,, \tag{3}$$

expressed as a fraction of the full image, where $\bar{\mathbf{s}} = (\bar{w}, \bar{h})$. We shift the crop inside the image when necessary, run the frozen Stage-1 and extent/joint-geometry modules on it, and map their predictions back to global coordinates. The *direct-crop control* uses this mapped prediction without learned fusion.

The learned refiner preserves the implementation's candidate-mixture structure without exposing training-specific parameterization. Let $\{\mathbf{c}_j\}$ denote candidate centers and $\{\mathbf{s}_k\}$ positive candidate extents. For variant $\nu \in \{\mathrm{geo}, \mathrm{vis}\}$, the refiner predicts normalized mixture weights together with residual parameters that are bounded through $\tanh$,

$$(\boldsymbol{\alpha}^\nu, \boldsymbol{\beta}^\nu, \mathbf{r}_c^\nu, \mathbf{r}_s^\nu) = F_{\eta_\nu}(\mathbf{u}^\nu), \qquad \alpha_j^\nu, \beta_k^\nu \geq 0, \quad \sum_j \alpha_j^\nu = 1, \quad \sum_k \beta_k^\nu = 1. \tag{4}$$

The refined geometry is

$$\begin{aligned} \hat{\mathbf{c}}^\nu &= \mathrm{clip}\left(\sum_j \alpha_j^\nu \mathbf{c}_j + \lambda_c\, \mathbf{s}_{\mathrm{crop}} \odot \tanh(\mathbf{r}_c^\nu),\ 0,\ 1\right), \\ \hat{\mathbf{s}}^\nu &= \mathrm{clip}\left(\exp\left[\sum_k \beta_k^\nu \log \mathbf{s}_k + \lambda_s \tanh(\mathbf{r}_s^\nu)\right],\ s_{\min},\ s_{\max}\right). \end{aligned} \tag{5}$$

Here, $\mathbf{s}_{\mathrm{crop}}$ is the normalized width and height of the realized crop; λ_c and λ_s are fixed residual scales; and $s_{\min}$ and $s_{\max}$ bound the extents. The operations $\log$, $\exp$, $\tanh$, and clip act elementwise, and $\odot$ denotes elementwise multiplication. The input $\mathbf{u}^{\mathrm{geo}}$ contains frozen full-image and crop-derived geometric signals. The input $\mathbf{u}^{\mathrm{vis}} = [\mathbf{u}^{\mathrm{geo}}; \mathbf{q}_{\mathrm{loc}}; \mathbf{h}_{\mathrm{loc}}]$ additionally concatenates frozen crop-query and local-context features. Thus, both variants use crop-derived geometry; the local-evidence variant also uses crop visual-language features. The base MLLM and Stage-1 components remain frozen while the repair components are fitted on the specified training split.

All learned diagnostic and repair components in the primary analysis are fitted on the DIOR training split. We use CounterGround-DIOR validation for diagnosis and the primary repair comparison;

Table 1: **Paired CounterGround benchmark.** Results are reported on CounterGround-COCO and the held-out CounterGround-DIOR test cohort. Sample and Pair are IoU ≥ 0.5 accuracies. Direction, Association, and ID✓/IoU× are evaluated over pairs with two valid predictions. All values are percentages. † denotes a supplementary specialist comparator rather than an MLLM.

Model	Sample	Pair	Direction	Association	ID✓/IoU×
CounterGround-COCO					
InternVL3.5-2B	41.10	20.84	95.00	72.73	51.55
MiniCPM-V-4.6	47.21	26.71	91.87	66.01	32.39
Qwen2.5-VL-3B-Instruct	64.77	44.91	87.44	68.89	23.85
Qwen3-VL-2B-Instruct	66.34	47.06	86.30	62.82	15.85
Qwen3.5-2B	64.14	45.01	86.40	60.08	15.07
Qwen3.5-4B	68.44	50.39	87.77	68.10	17.81
Qwen3.5-9B	70.50	52.84	91.19	72.11	19.28
GroundingDINO-base + resolver†	63.94	42.95	71.63	57.24	13.49
CounterGround-DIOR					
InternVL3.5-2B	29.45	15.63	98.47	82.24	66.26
MiniCPM-V-4.6	9.61	3.49	98.76	66.28	62.57
Qwen2.5-VL-3B-Instruct	35.70	22.35	95.49	65.03	42.68
Qwen3-VL-2B-Instruct	49.70	36.40	95.85	70.03	33.55
Qwen3.5-2B	50.48	37.26	95.23	73.20	35.97
Qwen3.5-4B	56.07	44.14	97.70	81.99	37.85
Qwen3.5-9B	57.33	45.81	97.51	82.21	36.38
GroundingDINO-base + resolver†	24.76	15.65	74.77	43.27	27.32

held-out test data are used neither for training nor for model selection. Learned repair conditions use fixed training schedules and predeclared seeds. The accompanying implementation specifies model revisions, optimization schedules, batch sizes, crop parameters, and reproducibility settings.

5 Experiments

Our experiments cover benchmark evaluation, controlled diagnosis, repair, and generalization. We evaluate seven MLLM configurations on the two CounterGround test sets, including Qwen3.5 at 2B, 4B, and 9B parameters. GroundingDINO-base with a relation resolver serves as a supplementary specialist comparator. Controlled Diagnosis uses Qwen3.5-2B and Qwen3-VL-2B-Instruct on CounterGround-DIOR validation; the primary repair study uses Qwen3.5-2B on the same set, with components fitted only on DIOR training data.

The DIOR-trained repair configuration is fixed before held-out evaluation. For CounterGround-COCO, we refit dataset-specific repair components on COCO training data. RefCOCO (Kazemzadeh et al., 2014; Yu et al., 2016) and VRSBench (Li et al., 2024a) use their native unpaired protocols, with components fitted on the corresponding training data. VRSBench evaluation includes 16,146 valid referring expressions from the official evaluation split, excluding 13 records with non-positive-area GT boxes. Localization accuracy uses an IoU threshold of 0.5 unless stated otherwise; VRSBench also reports Acc@0.7. Pair Accuracy confidence intervals use 10,000 bootstrap resamples of complete pairs with seed 42.

5.1 Paired Evaluation Reveals Behavioral Separation

Table 1 shows that relation-consistent responses and correct instance associations can coexist with localization failures across models and domains. For InternVL3.5-2B on CounterGround-DIOR, Direction Accuracy and Instance Association Accuracy reach 98.47% and 82.24% among valid pairs, while Pair Accuracy is 15.63% over all pairs. More directly, 66.26% of valid pairs associate both predicted centers with the intended instances but fail paired IoU. This last metric establishes the separation within the same evaluated pairs.

Table 2: **Controlled Diagnosis on CounterGround-DIOR validation** (2,298 pairs; 4,596 samples). GT-assisted conditions serve only as diagnostic interventions. Accuracies are percentages; mean IoU is on a 0-100 scale.

Model	Condition	Sample Acc.	Pair Acc.	Mean IoU
Qwen3.5-2B	Free-form grounding	59.70	48.65	55.49
	GT-candidate selection	**93.76**	**88.44**	**93.83**
	Fixed-center extent prediction	53.60	43.60	47.74
	GT-size oracle	71.89	62.01	59.67
Qwen3-VL-2B-Instruct	Free-form grounding	59.20	46.91	54.43
	GT-candidate selection	**84.12**	**72.08**	**84.28**
	Fixed-center extent prediction	54.33	44.71	47.71
	GT-size oracle	68.30	57.75	57.44

Table 3: **Diagnosis-guided repair on Qwen3.5-2B, CounterGround-DIOR validation.** Pair Accuracy intervals use 10,000 bootstrap resamples of complete pairs. Accuracies are percentages; mean IoU is on a 0-100 scale. Δ Pair is the percentage-point gain over free-form grounding.

Method	Sample Acc.	Pair Acc.	Pair 95% CI	Mean IoU	Δ Pair
Free-form grounding	59.70	48.65	[46.61, 50.65]	55.49	–
Center-only correction	54.91	44.60	[42.61, 46.60]	48.83	-4.05
Direct-crop control	28.92	18.45	[16.88, 20.02]	36.47	-30.20
Joint geometry refinement	67.96	56.06	[54.11, 58.06]	55.30	+7.41
Joint geometry + local evidence	**72.45**	**61.33**	**[59.38, 63.23]**	**58.03**	**+12.68**

On CounterGround-DIOR, scaling Qwen3.5 from 2B to 9B raises Sample Accuracy from 50.48% to 57.33%, Pair Accuracy from 37.26% to 45.81%, and Instance Association Accuracy from 73.20% to 82.21%; Direction Accuracy remains above 95%. Nevertheless, ID✓/IoU× is 36.38% at 9B, compared with 35.97% at 2B. On CounterGround-COCO, Pair Accuracy rises from 45.01% to 52.84% and Instance Association Accuracy from 60.08% to 72.11%, yet 19.28% of valid Qwen3.5-9B pairs have correct instance associations and fail paired IoU. Scaling improves absolute accuracy without eliminating the observed separation.

All 5,635 test pairs pass the eight construction checks in Section 3.2, confirming that this analysis uses structurally valid target-switch pairs.

5.2 Controlled Diagnosis: Explicit Target Selection Exceeds Free-Form Box Localization

GT-candidate selection yields substantially higher Pair Accuracy than free-form grounding for both models (Table 2): 88.44% vs. 48.65% for Qwen3.5-2B, and 72.08% vs. 46.91% for Qwen3-VL-2B-Instruct. These comparisons show that target selection with supplied GT candidates is easier than free-form localization; they do not establish equivalent target-selection ability without that assistance.

The fixed-center diagnostics further distinguish instance association from box precision. Independently estimated centers are associated with the intended instance for more than 94% of samples on both backbones, yet fixed-center extent prediction remains weaker than free-form grounding. Supplying GT size improves Pair Accuracy to 62.01% and 57.75%, respectively, but leaves substantial residual error. These results motivate joint geometric refinement beyond the tested fixed-center configurations.

5.3 Joint Geometry and Additional Local Features

Table 3 compares the repair conditions. The tested center-only configuration underperforms free-form grounding, whereas joint geometry refinement raises Pair Accuracy from 48.65% to 56.06%. Adding local visual-language features raises it to 61.33%, a total gain of 12.68 percentage points. The matched direct-crop control achieves only 18.45%. Mean IoU, however, decreases slightly with joint

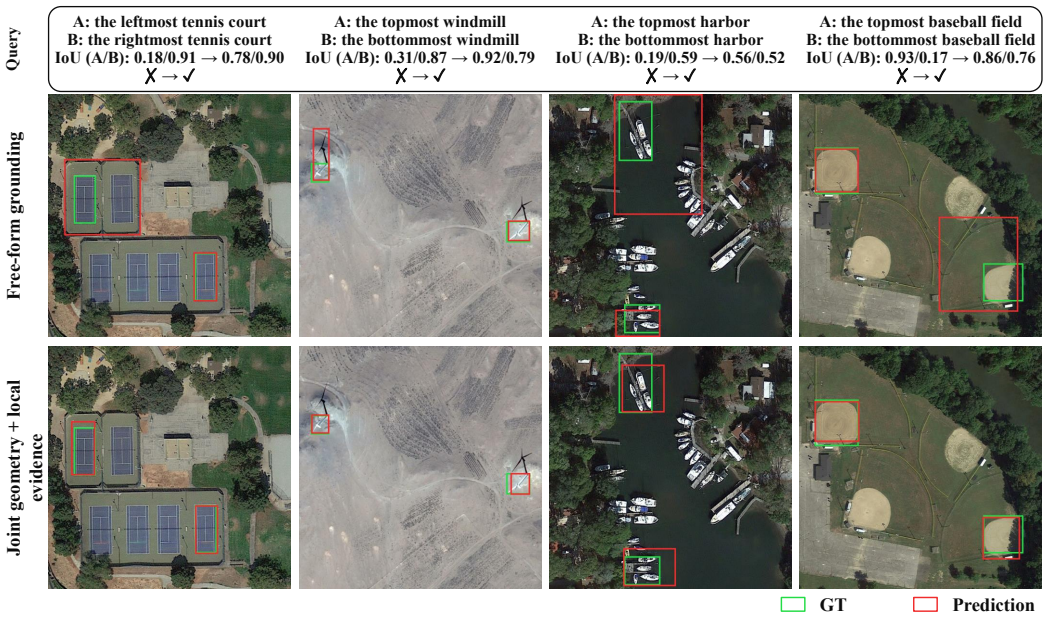

Figure 3: **Qualitative examples of diagnosis-guided repair.** Each column shows a CounterGround query pair. The top row uses free-form grounding, while the bottom row applies joint geometry refinement with fresh local visual evidence. Green boxes denote ground truth and red boxes denote predictions. These examples are illustrative; the aggregate repair results are reported in Table 3.

geometry alone (55.49 to 55.30 on a 0-100 scale) and improves to 58.03 with the full repair. Figure 3 illustrates paired predictions before and after the full repair.

5.4 Generalization Beyond the Primary Setting

We evaluate the repair formulation across the held-out, cross-backbone, cross-domain, and native grounding settings described above. Because components are refitted for each dataset, these experiments assess the applicability of the formulation across settings, rather than zero-shot transfer of one fitted model.

Full repair improves the Pair Accuracy point estimate in all four paired settings and the accuracy point estimate on every reported RefCOCO split and VRSBench metric (Table 4). Joint geometry alone does not improve every metric: VRSBench Acc@0.7 decreases from 28.64% to 27.62%, before increasing to 41.95% with full repair. These results support the repair formulation across the specified fitting protocols; the component diagnosis remains limited to CounterGround-DIOR validation.

6 Conclusion

CounterGround evaluates visual-grounding reliability through controlled target switches rather than independent query instances. Across two visual domains and multiple models, paired evaluation separates relation-consistent responses, intended-instance association, and precise localization. Scaling Qwen3.5 from 2B to 9B improves grounding accuracy without removing this separation. Controlled Diagnosis on two MLLMs shows that selection among explicit GT candidates can achieve substantially higher accuracy than free-form localization, while fixed-center geometric interventions leave residual error. Joint geometric refinement improves Pair Accuracy, and local visual-language features provide a further gain. The repair improves the reported point estimates on held-out CounterGround-DIOR with both backbones, on CounterGround-COCO, and under native RefCOCO and VRSBench protocols. These results support evaluating both target-switch behavior and box precision and show how paired diagnosis can guide repair.

Table 4: **Generalization of the repair formulation.** (a) Paired CounterGround evaluation across two benchmarks and two backbones. (b) Native unpaired evaluation on RefCOCO and VRSBench. Accuracies are percentages; gains are percentage-point differences computed from the displayed accuracies.

(a) Paired CounterGround evaluation

Backbone	Free-form		Repair		Δ Pair
	Sample Acc.	Pair Acc.	Sample Acc.	Pair Acc.	
CounterGround-DIOR					
Qwen3.5-2B	50.48	37.26	**60.97**	**47.80**	**+10.54**
Qwen3-VL-2B-Instruct	49.70	36.40	**63.17**	**49.97**	**+13.57**
CounterGround-COCO					
Qwen3.5-2B	64.14	45.01	**68.82**	**51.86**	**+6.85**
Qwen3-VL-2B-Instruct	66.34	47.06	**67.12**	**49.87**	**+2.81**

(b) Standard unpaired grounding benchmarks

Benchmark	Split / metric	Free-form grounding	Joint geometry	Full repair	Δ Full–Free
RefCOCO UNC	val Acc@0.5	88.04	92.70	**93.35**	**+5.31**
RefCOCO UNC	testA Acc@0.5	90.24	94.38	**95.05**	**+4.81**
RefCOCO UNC	testB Acc@0.5	83.91	89.81	**90.91**	**+7.00**
VRSBench	eval Acc@0.5	49.07	65.31	**73.00**	**+23.93**
VRSBench	eval Acc@0.7	28.64	27.62	**41.95**	**+13.31**

AI USE STATEMENT

We used generative AI tools, including ChatGPT and Codex, for language editing, code refactoring and debugging, literature searches, and discussions of experimental organization. The authors made all final decisions on benchmark design, experimental protocols, data and model selection, result interpretation, and scientific claims. We reviewed all AI-assisted work. Code changes were inspected and checked through repository validation and reruns; manuscript text, citations, equations, and reported results were checked against the implementation, experimental artifacts, and cited sources. The authors take responsibility for the final text, claims, code, and other artifacts produced with AI assistance.

REFERENCES

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. Cops-Ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. Ref-Adv: Exploring MLLM visual reasoning in referring expression tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

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 the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1780–1790, October 2021.

Weitai Kang, Jason Kuen, Mengwei Ren, Zijun Wei, Yan Yan, and Kangning Liu. VGent: Visual grounding via modular design for disentangling reasoning and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41160–41170, June 2026.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, 2014.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. DyFo: A training-free dynamic focus visual search for enhancing LMMs in fine-grained visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9098–9108, June 2025.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, June 2022.

Rang Li, Lei Li, Shuhuai Ren, Hao Tian, Shuhao Gu, Shicheng Li, Zihao Yue, Yudong Wang, Wenhan Ma, Zhe Yang, Jingyuan Ma, Zhifang Sui, and Fuli Luo. GroundingME: Exposing the visual grounding gap in MLLMs through multi-dimensional evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2412–2422, June 2026.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, YiQing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. GroundingGPT: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6657–6678, August 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755, 2014.

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 the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15440–15457, November 2024a.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. CLEVR-Ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024b.

Gabriele Lombardo, Luigi Maiorana, Liliana Lo Presti, and Marco La Cascia. Investigating anisotropy in visual grounding under controlled counterfactual perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4058–4067, June 2026.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024.

Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41171–41181, June 2026.

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

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-LMM: Grounding frozen large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24710–24721, June 2025.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. See say and segment: Teaching LMMs to overcome false premises. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13459–13469, June 2024.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Zhihan Yu and Ruifan Li. Revisiting counterfactual problems in referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13438–13448, June 2024.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. LLaVA-Grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pp. 19–35, 2024.

Peirong Zhang, Yidan Zhang, Luxiao Xu, Jinliang Lin, Zonghao Guo, Fengxiang Wang, Xue Yang, Kaiwen Wei, and Lei Wang. GeoViS: Geospatially rewarded visual search for remote sensing visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14335–14345, June 2026.