KNOWING HOW MANY IS NOT KNOWING WHICH ONES: NATIVE COMPLETE-SET GROUNDING IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized Referring Expression Comprehension (GREC) requires a model to recover all objects referred to by an expression, including no-target and multi-target cases. We identify a spurious-correlation failure mode in GREC: target cardinality remains partly predictable when visual evidence is removed or mismatched, revealing reliance on language and dataset priors. Yet accurate counting does not imply correct target selection. A strong query-based set decoder predicts the correct count on 98.38% of two-target expressions but achieves only 10.90% Pr for the complete target set. When the query changes to refer to a different pair of objects in the same image, the decoder usually preserves the count but rarely switches to the correct targets. We therefore propose **Native Complete-Set Grounding**, which trains the MLLM to select and localize all referred objects through its autoregressive localization pathway. In a matched comparison with a separate set-prediction decoder, Pr improves from 50.24% to 77.22%. Our 3B model achieves **81.24/76.18/68.11%** Pr on gRefCOCO val/testA/testB, establishing state-of-the-art performance on all three splits. The gains persist at 7B scale while conventional REC performance remains strong.

1 INTRODUCTION

Generalized Referring Expression Comprehension (GREC) extends conventional single-target grounding to no-target and multi-target expressions, making target cardinality an important part of the task (He et al., 2023). Yet we find that accurate cardinality can be a misleading indicator of successful grounding. A strong query-based set-prediction decoder predicts the correct target count on 98.38% of two-target expressions but recovers the complete referred set in only 10.90%. This discrepancy exposes a spurious-correlation failure mode in which a model can know how many targets to return without reliably identifying which objects should be grounded. Figure 1 summarizes this central finding and the main evidence motivating our study.

We further probe this failure mode through two complementary interventions. First, we construct same-image query pairs that preserve the ground-truth cardinality while changing the referred object set. The separate decoder predicts both target counts correctly in 97.59% of pairs, yet switches to the correct targets in only 4.02%. Cardinality can therefore remain highly stable even when the model fails to follow the changed referring expression. Second, when informative visual evidence is removed or mismatched, aggregate count accuracy remains above the majority baseline, indicating that target-count prediction can be partly supported by language and dataset priors. Together, these results show that apparently reliable cardinality can arise without faithful expression-conditioned target selection.

This distinction is critical for complete generalized grounding. Correct cardinality is necessary, but it is not sufficient: a prediction may return the right number of boxes while omitting a referred object, including an irrelevant one, or selecting the wrong members of the target set. Reliable GREC therefore requires recovering the complete referred set by determining not only how many targets should be returned, but also which objects satisfy the expression and where they are located. Similar failures have been observed in VLMs that rely on spurious image-text associations, visual-language priors, cross-modal correlations, or superficial textual patterns rather than task-relevant multimodal

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.

evidence (Varma et al., 2024; Luo et al., 2025; Chen et al., 2025a; 2026). Our results reveal an analogous failure in generalized grounding, where target cardinality can appear reliable despite incorrect target membership. The problem is not predicting target cardinality itself, but treating target count separately from deciding which objects are actually referred to.

We therefore propose Native Complete-Set Grounding, a simple autoregressive method for complete variable-cardinality grounding. The key insight is that complete-set recovery requires more than predicting the target cardinality; it requires determining which objects belong to the referred set while localizing their spatial positions. Therefore, instead of separating target prediction from the MLLM's grounding process, we directly train the MLLM to select and localize all referred objects through its autoregressive coordinate-generation pathway. By aligning complete-set supervision with the model's native localization interface, target selection and spatial localization are learned within the same autoregressive generation process. Each generated target is conditioned on the image, referring expression, and previously generated outputs, allowing the model to recover the complete referred set through a unified grounding process. The same interface handles no-target, single-target, and multi-target expressions without an auxiliary box-regression objective, bipartite-matching loss, or separate target-selection stage.

Native Complete-Set Grounding consistently improves complete-set recovery. In a matched comparison with a separate set-prediction decoder, official-val Pr improves from 50.24% to 77.22%, while directional switch accuracy on count-preserving query pairs increases from 4.02% to 65.68%. Our 3B model further achieves 81.24/76.18/68.11% Pr on gRefCOCO val/testA/testB, establishing state-of-the-art performance on all three official splits. The same formulation remains effective at 7B scale while conventional RefCOCO-family performance remains strong.

Our contributions are threefold:

1. We identify a spurious-correlation failure mode in generalized grounding, where apparently reliable target cardinality can coexist with incorrect target selection. Count-preserving query changes and visual-evidence interventions show that cardinality prediction can remain stable even when the model fails to follow the referred object set and can be partly supported by language and dataset priors.

2. We propose Native Complete-Set Grounding, which selects and localizes the complete referred set through the MLLM's native autoregressive localization pathway, without a separate set-prediction decoder or auxiliary target-selection objective.

3. We provide extensive empirical evidence through matched architectural comparison, diagnostic interventions, official gRefCOCO evaluation, 7B scaling, and conventional REC evaluation.

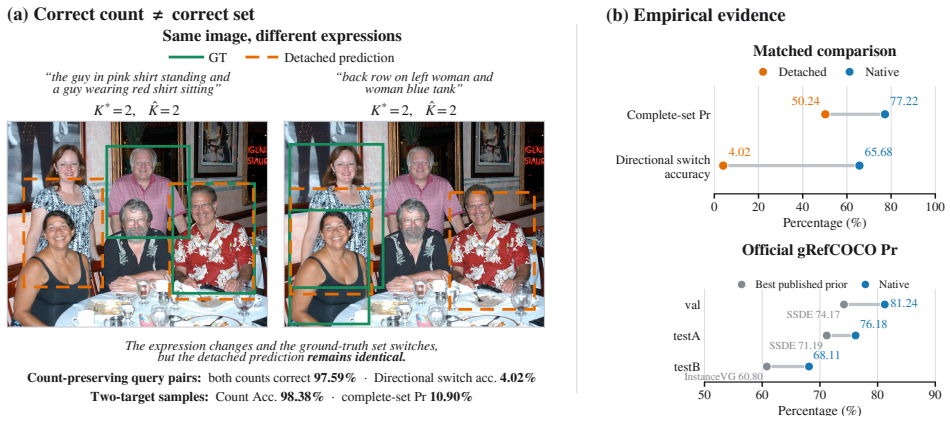

Figure 1: **Motivation and main evidence.** Correct cardinality can coexist with incorrect target selection: detached predictions remain stable under count-preserving query changes, while Native improves complete-set recovery, directional switching, and official gRefCOCO Pr.

2 Related Work

2.1 Referring Expression Comprehension and Generalized Visual Grounding

Standard REC benchmarks and methods study single-target grounding through region-language reasoning and end-to-end localization (Yu et al., 2016; Mao et al., 2016; Nagaraja et al., 2016; Yu et al., 2018; Chen et al., 2021; Kamath et al., 2021; Deng et al., 2021; Li et al., 2022; Dai et al., 2024; Wang et al., 2024b). GREC extends this setting to no-target and multi-target expressions (He et al., 2023; Ding et al., 2026), while GRES studies generalized referring at mask level (Liu et al., 2023).

Recent GREC methods address complementary aspects of variable-cardinality grounding: ReLA models region–region and region–language dependencies (Ding et al., 2026); HieA2G models adaptive target number (Wang et al., 2025); PropVG strengthens proposal generation and multi-granularity discrimination (Dai et al., 2025); InstanceVG develops instance-aware joint learning (Dai et al., 2026); and SSDE improves semantic-spatial discrimination and no-target recognition (Lei & Zhang, 2026). Multi-instance segmentation methods including GSVA, InstAlign, UniRES++, and InstructSeg study related generalized or multi-target settings (Xia et al., 2024; Nguyen et al., 2024; Liu et al., 2025; Wei et al., 2025). Our work asks whether accurate target cardinality is accompanied by faithful complete-set target selection and localization.

2.2 Grounding with Multimodal Large Language Models

Grounding-capable MLLMs provide generative spatial interfaces for region referring, coordinate generation, or unified localization (Wang et al., 2023; Peng et al., 2024; Chen et al., 2023; Bai et al., 2023; You et al., 2024; Xuan et al., 2024; Wang et al., 2024a; Xiao et al., 2024); Groma instead introduces localized region tokens during visual tokenization (Ma et al., 2024). These works show that pretrained MLLMs already expose spatial capabilities that can be adapted rather than replaced.

Other methods attach specialized spatial or pixel-level modules (Lai et al., 2024; Rasheed et al., 2024; Ren et al., 2024; Zhang et al., 2024; He et al., 2024; Xia et al., 2024), recover grounding from MLLM features or attention heads (Wu et al., 2025; Kang et al., 2025), separate reasoning from spatial prediction (Kang et al., 2026), or learn grounding tokens (Ren et al., 2026). Our focus is instead on where variable-cardinality target selection should be performed.

Spurious multimodal correlations are studied by RaVL and DIAL (Varma et al., 2024; Yalavarthi et al., 2026), ViLP (Luo et al., 2025), Cross-modal Causal Relation Alignment (Chen et al., 2025a), and Safety Mirage (Chen et al., 2026). We study the analogous case in generalized grounding, where cardinality can appear reliable despite incorrect target selection.

2.3 Autoregressive Localization and Set Prediction

Pix2Seq and UniTAB generate spatial outputs autoregressively (Chen et al., 2022; Yang et al., 2022), whereas DETR, DAB-DETR, and DINO use query-based set prediction and bipartite matching (Carion et al., 2020; Liu et al., 2022; Zhang et al., 2023). Permutation-invariant set modeling also includes Deep Sets, Set Transformer, latent-space set prediction, and cardinality-aware objectives (Zaheer et al., 2017; Lee et al., 2019; Preechakul et al., 2021; Cortes et al., 2024). Both paradigms are natural candidates for GREC; we compare them under a matched protocol and ask whether correct cardinality is accompanied by faithful target selection.

3 Method

Generalized Referring Expression Comprehension requires recovering the complete set of objects described by a referring expression, with target cardinality varying from zero to many. We build on the localization capability already available in grounding-capable multimodal large language models and train the model to generate the complete referred set directly through its autoregressive localization pathway. As illustrated in Figure 2, target selection and spatial localization are learned within the same generative grounding process rather than delegated to a separate set-prediction module.

3.1 Generalized Grounding as Complete-Set Prediction

Given an image I and a referring expression q, let the ground-truth target set be

$$B^* = \{b_1^*, \ldots, b_{K^*}^*\}, \qquad K^* = |B^*|, \tag{1}$$

where each target box is $b_i^* = (x_{1,i}, y_{1,i}, x_{2,i}, y_{2,i})$. The formulation covers no-target ($K^* = 0$), single-target ($K^* = 1$), and multi-target ($K^* \geq 2$) expressions. A model predicts $\hat{B} = f_\theta(I, q)$ with $\hat{K} = |\hat{B}|$. Correct cardinality is necessary for complete-set recovery but is not sufficient:

$$\hat{K} = K^* \not\Rightarrow \hat{B} \text{ correctly recovers } B^*. \tag{2}$$

GREC therefore requires the model to determine how many targets are referred to, which objects satisfy the expression, and where they are located. These requirements are conceptually distinct but need not be implemented as separate prediction heads. In our formulation, supervision contains all and only the referred objects, and a prediction is correct only when the complete target set is recovered without introducing irrelevant objects.

3.2 Native Complete-Set Grounding

A modular set predictor first computes multimodal representations $H = h_\theta(I, q)$ and delegates spatial prediction to an additional module g_ϕ, yielding $\hat{B} = g_\phi(H)$. Native Complete-Set Grounding instead predicts the complete referred set through the MLLM's autoregressive localization pathway:

$$\hat{y} \sim p_\theta(y \mid I, q), \qquad \hat{B} = \mathrm{Parse}(\hat{y}). \tag{3}$$

The same pretrained pathway that maps language-conditioned multimodal representations to spatial outputs is therefore adapted from individual-object localization to complete variable-cardinality grounding.

Let the generated response be $y = (y_1, \ldots, y_T)$. The MLLM factorizes its probability autoregressively as

$$p_\theta(y \mid I, q) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, I, q). \tag{4}$$

For a valid ordering $\pi = (\pi_1, \ldots, \pi_{K^*})$ of the ground-truth targets, complete-set generation can be viewed conceptually as

$$p_\theta\big(b_{\pi_1}^*, \ldots, b_{\pi_{K^*}}^* \mid I, q\big) = \prod_{j=1}^{K^*} p_\theta\Big(b_{\pi_j}^* \mid b_{\pi_{<j}}^*, I, q\Big). \tag{5}$$

This object-level factorization is conceptual. In implementation, each bounding box uses the model's native coordinate representation, and the model factorizes over the serialized token sequence.

The key distinction is where target selection is performed. A separate set predictor performs target selection and localization after obtaining the MLLM representations. Native Complete-Set Grounding keeps both operations within the language-conditioned autoregressive localization process. Each generated target is conditioned on the image, the referring expression, and previously generated outputs, so deciding which object to produce and predicting its spatial position are handled by the same generative pathway.

This design reuses the localization capability acquired during multimodal pretraining rather than replacing it with an additional spatial predictor. No separate target-selection head, query-based set decoder, or auxiliary box-regression objective is introduced.

3.3 Serializing an Unordered Target Set

GREC outputs an unordered target set, whereas an autoregressive model produces an ordered sequence. We therefore represent B^* using a valid serialization

$$y_\pi = s_\pi(B^*), \qquad \pi \sim q(\pi \mid B^*), \tag{6}$$

where different valid orderings correspond to the same underlying target set.

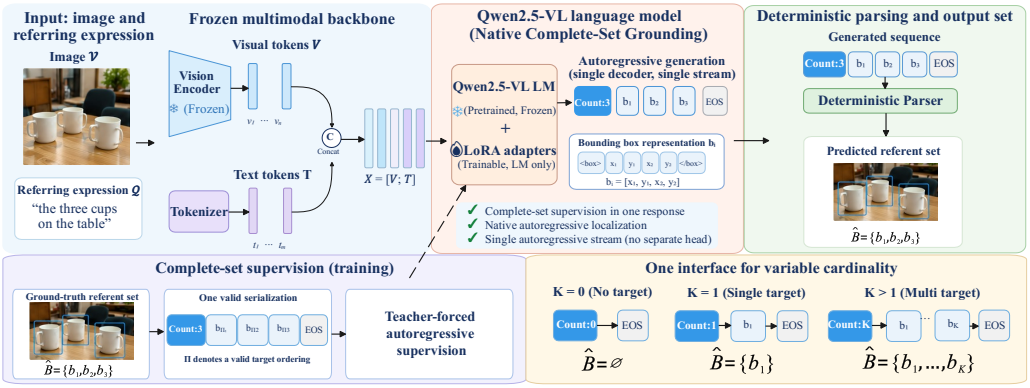

Figure 2: **Overview of Native Complete-Set Grounding.** Complete-set supervision uses one native autoregressive localization stream for variable-cardinality grounding, without a separate spatial decoder or target-selection head; exact serialization is given in Sec. 3.3 and Appendix A.

For the full-scale experiments, we use the structured Count+Boxes serialization as a unified output format for variable-cardinality grounding:

$$y_\pi = [\texttt{Count:}\quad K^*;\ \texttt{Boxes:}\quad b^*_{\pi_1}, \ldots, b^*_{\pi_{K^*}}], \tag{7}$$

with no-target examples represented by a zero count and an empty box list. For $K^* \geq 2$, the target ordering is varied reproducibly across epochs using the sample identifier, epoch index, and global seed. The referred objects and their cardinality remain unchanged; only their presentation order changes.

Importantly, neither the explicit count field nor target-order resampling defines Native Complete-Set Grounding. They are serialization choices for representing an unordered target set in an autoregressive response. The count is generated as part of the same sequence rather than predicted by an independent count head, while order resampling changes only the serialization of the same supervision. Alternative serialization choices are examined separately in Sec. 4.5.

3.4 Learning and Inference

Training uses the standard causal language-modeling objective over assistant-response positions $\mathcal{A}$:

$$\mathcal{L}_{\text{NCSG}} = -\mathbb{E}_{\pi \sim q(\pi|B^*)} \left[\sum_{t \in \mathcal{A}} \log p_\theta \left([y_\pi]_t \mid [y_\pi]_{<t}, I, q \right) \right]. \tag{8}$$

No additional count-classification loss, box-regression loss, bipartite-matching objective, or target-selection loss is introduced. The vision encoder and multimodal projector remain frozen, while the language-model pathway of Qwen2.5-VL (Bai et al., 2025) is adapted with LoRA (Hu et al., 2022).

At inference time, the model performs a single autoregressive generation, $\hat{y} = \text{GreedyDecode}_\theta(I, q)$, followed by $\hat{B} = \text{Parse}(\hat{y})$. A deterministic parser extracts the generated bounding-box representations, maps them to the evaluation coordinate system, and returns the final prediction set. A response is structurally valid only when its parsed cardinality and box list are consistent and all parsed boxes are spatially valid. Malformed or inconsistent generations remain part of evaluation rather than being repaired or discarded. Generation terminates at EOS or the fixed generation limit. No external detector, reranker, hard cardinality decoder, or auxiliary verifier is used.

The same inference interface therefore handles no-target, single-target, and multi-target expressions. Variable cardinality is represented directly by the generated response, while complete-set supervision trains the model to determine both which objects should be returned and where they are located.

4 EXPERIMENTS

We evaluate Native Complete-Set Grounding from five complementary perspectives: official gRefCOCO benchmarking, matched comparison with a separate set-prediction decoder, diagnostic interventions on target counting and selection, output-representation controls, and generalization across model scale and conventional REC benchmarks. We further analyze performance as target cardinality increases and report inference efficiency.

4.1 EXPERIMENTAL SETUP

Datasets and protocols. We use gRefCOCO (He et al., 2023) as the primary benchmark for generalized referring expression comprehension. Training uses the official training split, while model selection is performed exclusively on the official validation split; testA and testB are used only for final evaluation. RefCOCO, RefCOCO+, and RefCOCOg (Yu et al., 2016; Mao et al., 2016; Nagaraja et al., 2016) are used to evaluate compatibility with conventional single-target REC. Controlled comparisons use a fixed 12k training subset and a disjoint 3k development subset with no sample or image overlap with the controlled training data or official evaluation splits. The Native and separate-decoder formulations share the same Qwen2.5-VL-3B backbone, frozen visual components, LM-side LoRA adaptation, training data, seed, and training budget.

Metrics. All paper-facing metrics are computed from the final predicted set $\hat{B}$. We report complete-set Pr, Set F1@0.5 under one-to-one IoU matching, target-count accuracy, and Pr conditioned on correct target count. N-acc measures correct rejection on ground-truth no-target expressions; malformed or structurally invalid outputs remain in the denominator and are counted as incorrect.

For the diagnostic analyses below, we additionally report pair-level measures on same-image query pairs with identical target cardinality but different referred object sets. These include whether both target counts are correct, whether both complete target sets are recovered, whether the prediction switches to the targets specified by the changed expression, and whether predictions for the paired queries overlap.

4.2 MAIN RESULTS ON GREFCOCO

Table 1: **Comparison with recent methods on the official gRefCOCO benchmark (%).**

		val		testA		testB	
Method	Venue	Pr	N-acc.	Pr	N-acc.	Pr	N-acc.
ReLA (Ding et al., 2026)	IJCV 2026	61.90	56.37	50.35	59.02	44.61	58.40
HieA2G (Wang et al., 2025)	AAAI 2025	67.80	60.30	66.00	60.10	56.50	56.00
PropVG (Dai et al., 2025)	ICCV 2025	72.20	72.80	68.80	69.90	59.00	65.00
InstanceVG (Dai et al., 2026)	TPAMI 2026	73.50	72.80	70.20	71.10	60.80	65.20
SSDE (Lei & Zhang, 2026)	ACM MM 2026	74.17	75.93	71.19	72.93	60.62	67.28
Native Complete-Set Grounding (3B)	Ours	**81.24**	**80.39**	**76.18**	**75.13**	**68.11**	**70.38**

As shown in Table 1, Native Complete-Set Grounding achieves 81.24/76.18/68.11% Pr on gRefCOCO val/testA/testB, establishing state-of-the-art performance on all three official splits. N-acc reaches 80.39/75.13/70.38%, showing that the gain in complete-set recovery is accompanied by strong no-target rejection.

On val, the model additionally reaches 84.19% Set F1, 87.04% Count Acc., and 93.33% Pr given correct count. The high conditional Pr indicates that correct cardinality is often accompanied by successful recovery on this split, while the cardinality-wise analysis in Sec. 4.6 shows that this relationship weakens substantially as the number of targets grows.

4.3 TARGET COUNTING AND TARGET SELECTION

A central question in this work is whether accurate target counting reflects correct target selection. We compare Native Complete-Set Grounding with a separate query-based set predictor under a

matched protocol; Table 2 reports the architectural comparison and count-preserving query-pair diagnostics. The separate formulation applies a 12-query, three-layer DETR-style decoder to the MLLM representations and predicts objectness, bounding boxes, and a global target count. Spatial supervision uses Hungarian matching with objectness, L1, and GIoU terms.

Table 2: **Matched Native-versus-separate-decoder comparison (%).**

(a) Overall official-val results

Formulation	Pr	F1	Cnt. Pr|Cnt.
Separate decoder	50.24	58.22	82.80 60.67
Native	**77.22**	**81.45**	**86.27 89.52**

(b) Complete-set Pr

GT K	Sep.	Native
$K = 2$	10.90	**74.13**
$K \geq 3$	0.00	**35.71**

(c) Count-preserving query-pair analysis

Metric	Sep.	Native
Both counts correct	**97.59**	95.44
Pairwise complete set	0.00	**29.49**
Directional switch	4.02	**65.68**
Prediction overlap	86.86	**24.66**

Native improves official-val Pr from 50.24% to 77.22%, while Set F1 rises from 58.22% to 81.45%. The distinction between counting and target selection is particularly pronounced on two-target expressions: the separate decoder predicts the correct number of targets on 98.38% of samples but recovers the complete target set in only 10.90%.

We further isolate target selection by constructing same-image query pairs with the same ground-truth cardinality but different referred object sets. The separate decoder predicts both counts correctly in 97.59% of pairs, yet achieves only 4.02% directional switch accuracy and produces overlapping predictions for 86.86% of paired queries. Native preserves similarly high count accuracy while increasing directional switch accuracy to 65.68% and reducing prediction overlap to 24.66%. All major between-model differences are significant under image-cluster bootstrap confidence intervals.

These results show that accurate cardinality can be learned without reliable adaptation to the objects specified by the referring expression. Native grounding substantially strengthens expression-conditioned target selection rather than merely improving target counting.

4.4 Visual-Evidence Ablation

We next examine whether target-count prediction is fully grounded in the visual input. Table 3 evaluates the same model under three input conditions: the original image, a mean image that removes informative visual content, and a mismatched image paired with the original referring expression.

Table 3: **Target-count accuracy under visual-evidence ablation (%).**

Subset	Normal	Mean image	Mismatched	Majority
All samples	**84.05**	51.89	50.50	36.42
Implicit-cardinality	**84.30**	51.79	51.45	38.64
Positive-target	**88.38**	25.12	27.51	–

Removing or mismatching informative visual evidence reduces Count Acc. substantially, but aggregate accuracy remains around 50%, well above the 36.42% majority baseline. The same pattern holds for expressions without explicit cardinality words, where Mean-image and Mismatched-image accuracy remain at 51.79% and 51.45%, respectively.

This residual performance should not be interpreted as evidence that positive-target counting can be performed without vision. On positive-target expressions, Count Acc. drops from 88.38% to 25.12%/27.51% under Mean-image and Mismatched-image inputs. The degraded model also shows a strong tendency toward predicting the empty set; for example, on val with the mean image, Count Acc. is 97.89% for $K = 0$ but only 46.62% for $K = 2$.

The results therefore indicate that aggregate target-count accuracy mixes grounded visual counting with language and dataset priors, particularly the no-target prior. Visual evidence remains essential for reliable positive-target cardinality. Together with the count-preserving query-pair analysis in Sec. 4.3, this provides the empirical basis for the spurious-correlation failure mode described in Sec. 1.

4.5 Controlled Analysis of Serialization and Cardinality

Table 4 tests whether explicit cardinality, target-order resampling, or hard cardinality constraints account for the gains of native complete-set grounding under the controlled 12k/3k setting; these controls isolate potential sources of the gain rather than select the full-scale serialization.

Within this controlled setting, free-form autoregressive generation obtains the strongest result among the serialization variants. Explicit cardinality does not improve complete-set recovery, and target-order resampling provides no gain over fixed-order supervision. Hard cardinality enforcement changes Pr only marginally, from 77.22% to 77.25%, and leaves $K \geq 3$ Pr unchanged.

Table 4: **Serialization and cardinality controls under the matched 12k/3k protocol (%).**

(a) Training representation

Representation	Pr	Set F1	Count Acc.	Pr|Count
Free-form AR	**81.85**	**85.93**	**89.96**	**90.99**
Explicit K + fixed	78.89	83.02	87.43	90.23
Explicit K + resampling	77.22	81.45	86.27	89.52

(b) Cardinality-constrained decoding

Decoding	Pr	Count-box consistency	$K \geq 3$ Pr
Soft-K	77.22	99.92	35.71
Hard-K	77.25	99.97	35.71

Soft-K and Hard-K are nearly identical: enforcing count-box consistency from 99.92% to 99.97% changes Pr by only 0.03 points and leaves $K \geq 3$ Pr at 35.71%. These controls confirm that the effectiveness of Native Complete-Set Grounding does not arise from an explicit count token, target-order resampling, or hard cardinality constraints. The key distinction is the use of the MLLM's autoregressive localization pathway for complete-set target selection and localization.

4.6 Cardinality and Failure Analysis

Extending the count–selection discrepancy in Sec. 4.3, we examine how this relationship changes as the target set grows. Figure 3 summarizes cardinality-wise performance and the composition of high-K failures.

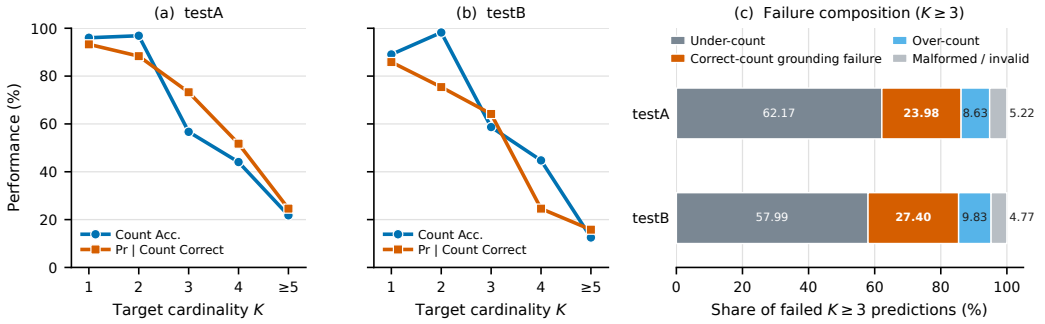

Figure 3: **Cardinality-wise performance and high-K failures.** Count Acc. and Pr|Count both decline as K grows, while a substantial fraction of $K \geq 3$ errors persists even when cardinality is correct.

On testA, Count Acc. remains above 96% for $K = 1$ and $K = 2$, but falls to 56.70%, 44.06%, and 21.79% for $K = 3, 4, \geq 5$; testB shows the same pattern. Correct cardinality does not remove the difficulty: Pr conditioned on correct count falls to 73.25/51.74/24.59% on testA and 64.17/24.58/15.79% on testB for $K = 3/4/ \geq 5$. Thus, high-cardinality degradation cannot be explained by counting errors alone.

The failure composition in Figure 3(c) further separates the two bottlenecks. Under-counting accounts for 62.17%/57.99% of failed $K \geq 3$ predictions on testA/testB, while 23.98%/27.40% fail despite predicting the correct cardinality.

Among correct-count grounding failures, 96.44% on testA and 98.68% on testB still contain at least one correctly grounded target, indicating partial-set recovery rather than complete loss of referent understanding. High-cardinality grounding is therefore limited by both target-count estimation and complete-set membership recovery. Figure 4 qualitatively illustrates the same distinction.

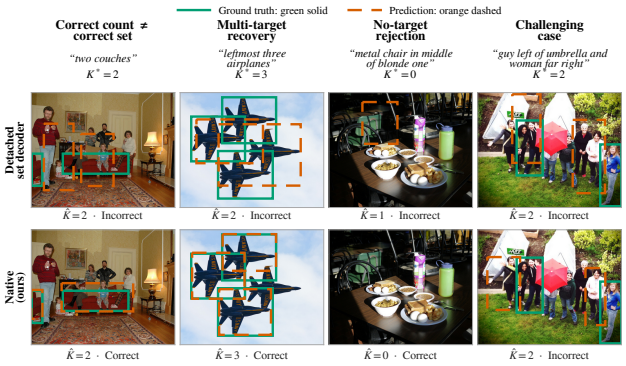

Figure 4: **Qualitative comparison of the separate decoder and Native.** Examples include correct-count selection errors, multi-target recovery, no-target rejection, and a shared hard case.

The two trends expose distinct failure sources as the target set grows: counting errors dominate the overall high-K failures, while membership errors remain substantial even after conditioning on the correct count. High-cardinality degradation therefore extends the count–selection gap rather than reducing to a counting problem alone.

4.7 Scaling and Conventional REC

We evaluate scale robustness and conventional REC compatibility. At 7B scale, Native Complete-Set Grounding improves Pr from 72.33% to 78.35% and N-acc from 79.48% to 82.01%, showing that the benefit is not specific to the 3B backbone. Table 5 reports conventional REC performance.

Table 5: **Recent conventional REC comparison using Acc@0.5 (%).**

	RefCOCO			RefCOCO+			RefCOCOg	
Method	val	testA	testB	val	testA	testB	val	test
TCRT'25 (Chen et al., 2025b)	**91.07**	92.85	86.24	84.55	90.26	76.59	84.41	85.13
GETok-SFT'26 (Ren et al., 2026)	90.60	**93.70**	**87.20**	**86.70**	**90.90**	**79.90**	**88.50**	**88.40**
Native Complete-Set Grounding (3B)	89.63	92.03	85.24	82.13	87.86	73.63	82.95	83.63

The adapted 3B model retains strong performance across all eight RefCOCO-family splits in Table 5, indicating that complete variable-cardinality adaptation remains compatible with conventional single-target REC.

4.8 Efficiency and Limitations

On 500 aligned official-testA examples (single RTX 4090, batch size 1, BF16), Native is 14.27× slower than the separate decoder with comparable peak memory, and latency rises as more targets are generated. The main limitations are high-K under-counting and incomplete set recovery despite correct cardinality, together with autoregressive cost that grows with output length. Improving coverage and decoding efficiency remain important directions.

5 Conclusion

Accurate cardinality does not guarantee faithful complete-set grounding. Count-preserving query changes expose target-selection failures, while visual-evidence ablations show partial reliance on language and dataset priors. We identify this spurious-correlation failure mode and propose Native Complete-Set Grounding, which unifies target selection and localization in the MLLM's autoregressive pathway. Native improves complete-set Pr from 50.24% to 77.22% and directional switch accuracy from 4.02% to 65.68%; the 3B model reaches 81.24/76.18/68.11% Pr on gRefCOCO val/testA/testB, and gains persist at 7B scale. High-K grounding and autoregressive latency remain key limitations.

Reproducibility Statement

The main paper specifies the task formulation, native autoregressive objective, evaluator definitions, model-selection protocol, and controlled comparison protocol. The appendix provides implementation, output-format, matched-comparison, diagnostic, cardinality, efficiency, and data/evaluator details. Training and evaluation code, configurations, and diagnostic outputs will be included in the anonymous supplementary material.

AI Use Statement

Generative AI tools were used only to assist with literature retrieval and discovery and to aid or polish the writing of this manuscript. All retrieved references and AI-assisted edits were manually checked by the authors against the original sources and the underlying experimental record. Generative AI tools were not used to develop the method, design or conduct experiments, generate data or results, or interpret experimental results. The authors take responsibility for the final content, claims, citations, and artifacts.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

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.

Long Chen, Wenbo Ma, Jun Xiao, Hanwang Zhang, and Shih-Fu Chang. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *AAAI*, 2021.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022.

Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsen Zheng, and Liang Lin. Cross-modal causal relation alignment for video question grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24087–24096, 2025a.

Wenbo Chen, Zhen Xu, Ruotao Xu, Si Wu, and Hau-San Wong. Task-aware cross-modal feature refinement transformer with large language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3931–3941, 2025b.

Yiwei Chen, Yuguang Yao, Yihua Zhang, Bingquan Shen, Gaowen Liu, and Sijia Liu. Safety mirage: How spurious correlations undermine vlm safety fine-tuning and can be mitigated by machine unlearning. In *International Conference on Learning Representations*, 2026.

Corinna Cortes, Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Cardinality-aware set prediction and top-k classification. In *NeurIPS*, 2024.

Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. In *NeurIPS*, 2024.

Ming Dai, Wenxuan Cheng, Jiedong Zhuang, Jiang-jiang Liu, Hongshen Zhao, Zhenhua Feng, and Wankou Yang. Propvg: End-to-end proposal-driven visual grounding with multi-granularity discrimination. In *ICCV*, 2025.

Ming Dai, Wenxuan Cheng, Jiang-Jiang Liu, Lingfeng Yang, Zhenhua Feng, Wankou Yang, and Jingdong Wang. Improving generalized visual grounding with instance-aware joint learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *ICCV*, 2021.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Yu-Gang Jiang. Grex: Generalized referring expression segmentation, comprehension, and generation. *International Journal of Computer Vision*, 134(2):79, 2026.

Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal instruction tuned llms with fine-grained visual perception. In *CVPR*, 2024.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr – modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *CVPR*, 2025.

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 *CVPR*, 2026.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019.

Kaiyan Lei and Xu-Yao Zhang. Semantic-spatial discriminability enhancement for generalized visual grounding. *arXiv preprint arXiv:2608.30233*, 2026. Accepted at ACM Multimedia 2026.

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 *CVPR*, 2022.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, 2023.

Jing Liu, Wenxuan Wang, Yisi Zhang, Yepeng Tang, Xingjian He, Longteng Guo, Tongtian Yue, and Xinlong Wang. Towards unified referring expression segmentation across omni-level visual target granularities. *arXiv preprint arXiv:2504.01954*, 2025.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *ICLR*, 2022.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in vlms. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156, 2025.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *ECCV*, pp. 417–435, 2024.

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

Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016.

E-Ro Nguyen, Hieu Le, Dimitris Samaras, and Michael S. Ryoo. Instance-aware generalized referring expression segmentation. *arXiv preprint arXiv:2411.15087*, 2024.

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 *ICLR*, 2024.

Konpat Preechakul, Chawan Piansaddhayanon, Burin Naowarat, Tirasan Khandhawit, Sira Sriswasdi, and Ekapol Chuangsuwanich. Set prediction in the latent space. In *NeurIPS*, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024.

Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. In *CVPR*, 2026.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *CVPR*, 2024.

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*, 2023.

Yabing Wang, Zhuotao Tian, Qingpei Guo, Zheng Qin, Sanping Zhou, Ming Yang, and Le Wang. Referencing where to focus: Improving visual grounding with referential query. In *NeurIPS*, 2024b.

Yaxian Wang, Henghui Ding, Shuting He, Xudong Jiang, Bifan Wei, and Jun Liu. Hierarchical alignment-enhanced adaptive grounding network for generalized referring expression comprehension. In *AAAI*, 2025.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *ICCV*, 2025.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. In *CVPR*, 2025.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, 2024.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024.

Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *CVPR*, 2024.

Bharat Chandra Yalavarthi, Nalini Ratha, and Venu Govindaraju. Label-free mitigation of spurious correlations in vlms using sparse autoencoders. In *International Conference on Learning Representations*, 2026.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *ECCV*, 2022.

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 *ICLR*, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *NeurIPS*, 2017.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*, 2024.

A Implementation and Output Protocol

Model adaptation. The main model is Qwen2.5-VL-3B-Instruct; scale validation uses Qwen2.5-VL-7B-Instruct (Bai et al., 2025). The vision encoder and multimodal projector are frozen. LoRA is applied to the language-model `q_proj`, `k_proj`, `v_proj`, and `o_proj` layers with rank 16, scaling factor 32, and dropout 0.05. Training uses AdamW, learning rate 2×10^{-5}, weight decay 0.01, BF16 precision, per-device batch size 1, eight-step gradient accumulation, and a three-epoch schedule. Images use the Qwen2.5-VL dynamic pixel budget of 12,544–451,584 pixels. Native inference uses greedy decoding with temperature 0, one beam, and at most 128 new tokens for the main configuration.

Assistant-only supervision. The native runs construct the multimodal prompt with the processor before masking. Prompt, image-processing, and padding positions are assigned label -100, so the causal objective is applied only to assistant-response tokens. This avoids estimating the assistant boundary from raw text length in a sequence that contains processor-inserted vision tokens.

Response format and parsing. The main response format is `Count: K` followed by `Boxes: [[x1,y1,x2,y2], ...]`; no-target examples use `Count: 0` and an empty box list. Coordinates are integer pixels in the processor-resized image and are mapped back to normalized coordinates for evaluation. N-acc is computed on GT-empty examples from the canonical final prediction set. A rejection is correct only for a structurally valid response whose final parsed box set is empty; malformed or inconsistent generations remain in the denominator and are counted as incorrect. Free-form controlled training omits the Count line and uses fixed manifest order; explicit-K controls use either fixed order or deterministic order resampling, as summarized in Table 6.

Table 6: **Output representations used in controlled and main training.**

Configuration	Count field	Target order
Free-form AR	No	Fixed manifest
Explicit K + fixed	Yes	Fixed manifest
Explicit K + resampling	Yes	Reproducibly varied
Main 3B model	Yes	Reproducibly varied

Controlled-study scope. The serialization and cardinality controls use the matched 12k/3k protocol and are analyzed independently from the full-scale experiments.

B Matched Architecture and Query-Pair Diagnostics

The controlled architecture comparison uses the same backbone adaptation, data split, seed, and training budget while changing only the spatial prediction pathway. Table 7 summarizes the design.

For Hungarian matching, the separate decoder uses objectness, L1, and generalized-IoU costs with weights 1, 5, and 2. Its training objective uses count, objectness, L1, and GIoU weights 1, 1, 5, and 2, with no-object classification weight 0.1. At inference, the count head predicts $\hat{K}$ and the top-$\hat{K}$ object queries by objectness are returned. The +26.99-point official-val Pr advantage of Native over the separate decoder has a paired sample-level bootstrap 95% CI of $[+26.22, +27.75]$ using 10,000 resamples (seed 42).

B.1 Count-Preserving Query-Pair Analysis

To isolate expression-conditioned target selection from cardinality, we construct same-image query pairs that share the same ground-truth target count but refer to different object sets. This design holds the image and cardinality fixed while changing which objects should be selected. We report four pair-level measures: whether both counts are correct, whether both complete target sets are correct, whether predictions switch to the targets specified by the changed expression, and prediction overlap across the paired queries.

Table 7: **Matched protocol for the controlled architecture comparison.**

Component	Native Complete-Set Grounding	Separate set-prediction decoder
Backbone	Qwen2.5-VL-3B-Instruct	Same
Train / held-out split	12k / 3k	Same
Random seed	42	42
Vision encoder / projector	Frozen / frozen	Frozen / frozen
LM adaptation	LoRA	LoRA
LoRA r/α/dropout	16 / 32 / 0.05	16 / 32 / 0.05
LM learning rate	2×10^{-5}	2×10^{-5}
Spatial predictor	Native AR localization	3-layer DETR-style decoder
Decoder width / heads / queries	–	512 / 8 / 12
Decoder FFN / dropout	–	2048 / 0.1
Decoder learning rate	–	1×10^{-4}
Cardinality prediction	Generated response	Explicit 0–12 count head
Matching	–	Hungarian
Training loss	Causal LM	Count + objectness + L1 + GIoU
Model selection	Held-out Pr	Held-out Pr
Paper evaluation	official val	official val

Table 8: **Count-preserving query-pair diagnostics (%).**

Metric	Separate decoder	Native
Both queries have correct count	**97.59**	95.44
Pairwise complete-set accuracy	0.00	**29.49**
Directional switch accuracy	4.02	**65.68**
Prediction overlap across paired queries	86.86	**24.66**

Uncertainty is estimated by clustering bootstrap resampling at the image level so that query pairs from the same image remain together. The major between-model differences reported in the main paper have confidence intervals that do not cross zero. The diagnostic therefore supports a target-selection interpretation rather than a change in target cardinality alone.

C Visual-Evidence Ablation Details

The visual-evidence ablation evaluates the same trained model with the original image, a mean-image replacement that removes informative object-level visual content, or a mismatched image paired with the original expression. The manipulation changes visual evidence at inference time without retraining the model. Table 9 reports split-wise Count Acc.; Table 10 reports the pooled subsets used in the main paper.

Table 9: **Split-wise Count Acc. under visual-evidence ablation (%).**

Split	Normal	Mean image	Mismatched image	Majority baseline
val	87.05	78.64	65.51	62.58
testA	84.33	40.10	53.60	30.94
testB	81.06	42.29	33.51	35.15
Pooled	**84.05**	51.89	50.50	36.42

Table 10: **Pooled subgroup analysis for visual-evidence ablation (Count Acc., %).**

Subset	Normal	Mean image	Mismatched image	Majority baseline
All samples	84.05	51.89	50.50	36.42
Implicit-cardinality expressions	84.30	51.79	51.45	38.64
Positive-target expressions	88.38	25.12	27.51	–

The split-wise results reveal why aggregate Count Acc. can remain deceptively high after visual information is degraded. In particular, no-target examples are substantially easier to classify under the mean-image condition; on val, Count Acc. is 97.89% for $K = 0$ but only 46.62% for $K = 2$. By contrast, positive-only Count Acc. collapses to 25.12%/27.51% under mean/mismatched inputs. Thus the ablation does not imply that target counting is vision-independent. Instead, aggregate cardinality accuracy mixes visually grounded evidence with expression and dataset priors, especially the empty-set prior. The intervention diagnoses the contribution of non-visual priors but does not identify a unique causal feature or shortcut.

D Additional Cardinality and Failure Analysis

Tables 11 and 12 provide the full cardinality breakdown underlying Figure 3. The conditional metric Pr|Count quantifies complete-set correctness among examples for which the predicted cardinality is already correct.

Table 11: **Cardinality-wise results on official testA (%).**

GT K	N	Pr	Set F1	Count Acc.	Pr|Count
1	5,917	89.57	90.10	96.03	93.28
2	5,940	85.57	92.78	96.87	88.34
3	1,813	41.53	77.01	56.70	73.25
4	522	22.80	67.36	44.06	51.74
≥ 5	560	5.36	42.86	21.79	24.59

Table 12: **Cardinality-wise results on official testB (%).**

GT K	N	Pr	Set F1	Count Acc.	Pr|Count
1	5,646	76.46	76.66	89.05	85.86
2	3,957	74.07	85.35	98.23	75.41
3	932	37.66	73.39	58.69	64.17
4	400	11.00	55.85	44.75	24.58
≥ 5	455	1.98	35.90	12.53	15.79

High-cardinality taxonomy. Among failed $K \geq 3$ examples, under-counting accounts for 62.17% of testA failures and 57.99% of testB failures. Correct-count grounding failures account for 23.98% and 27.40%, respectively; over-counting accounts for 8.63% and 9.83%, while malformed/invalid predictions account for 5.22% and 4.77%. Within correct-count grounding failures, 96.44% on testA and 98.68% on testB contain at least one correctly grounded member, indicating partial-set recovery rather than complete failure to identify any referent.

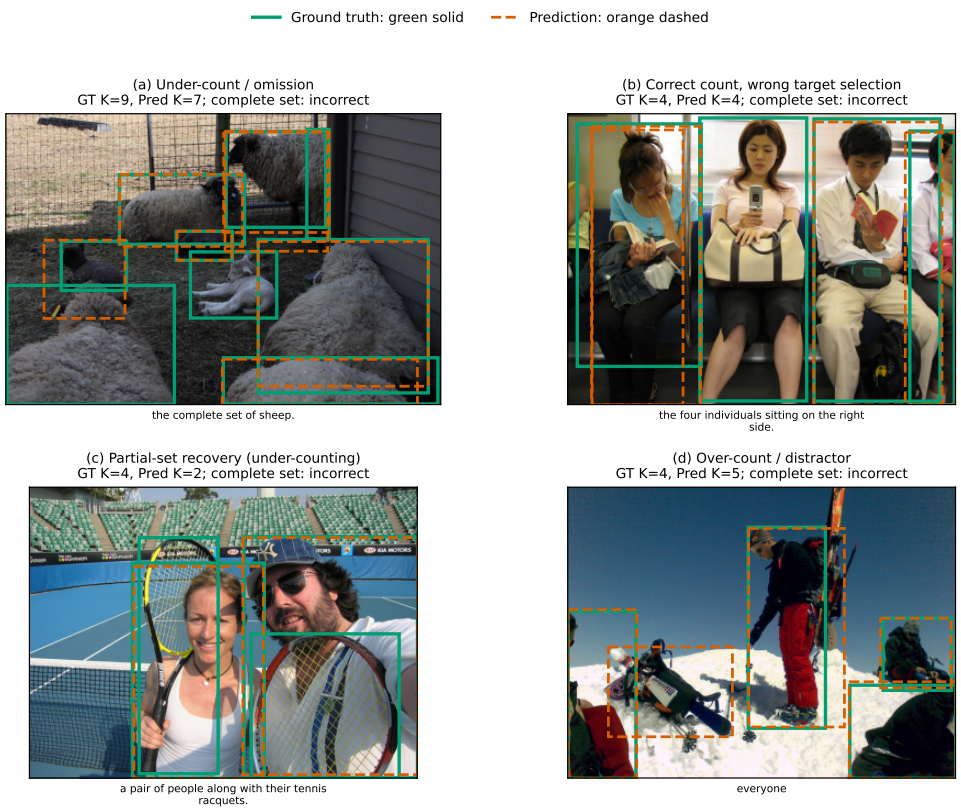

Figure 5: **Representative high-cardinality failure modes.** Examples illustrate under-counting, correct-count target-selection errors, partial-set recovery, and over-counting. Green boxes denote ground truth and orange boxes denote predictions.

E Efficiency Analysis

Efficiency is measured on 500 aligned official-testA samples using an otherwise idle RTX 4090, batch size 1, BF16, 10 warm-up samples, and CUDA-synchronized timing. Both methods process the same sample order sequentially. Timing covers the synchronized model forward/generation call; host-side preprocessing and deterministic parsing are excluded.

Table 13: **Isolated-GPU inference efficiency.**

Method	Mean ms	Median ms	Peak MiB	Tokens/sample
Native	1390.36	1257.84	7304.5	28.29
Separate decoder	97.46	95.49	7600.1	–

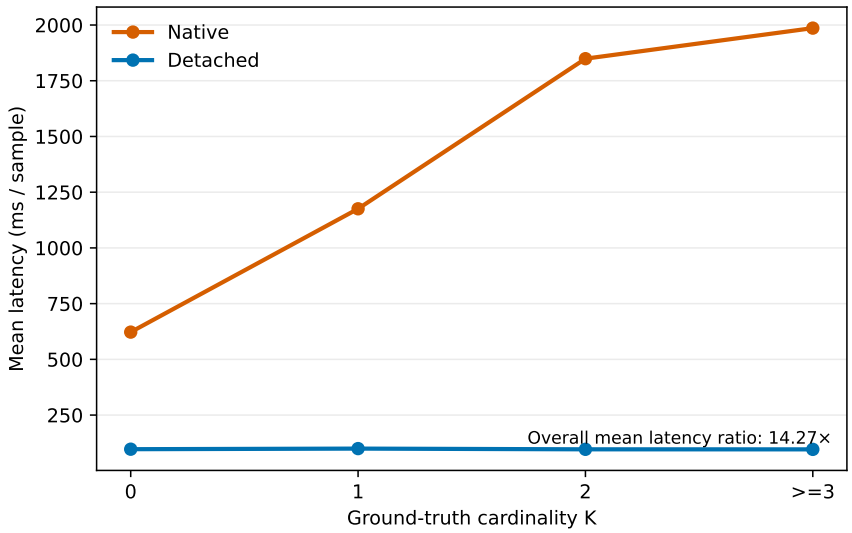

Figure 6: **Inference latency by ground-truth cardinality.** Native autoregressive latency rises with the number of target boxes, whereas separate parallel decoding remains approximately constant.

F Scaling and Conventional REC Details

The 7B scale comparison uses the same GREC evaluator as the main 3B study. The unadapted Qwen2.5-VL-7B model obtains 72.33 Pr and 79.48 N-acc. on official val, while Native complete-set adaptation obtains 78.35 Pr and 82.01 N-acc. No conventional REC test split participates in model selection.

Table 14: **Detailed conventional REC metrics for the adapted 3B model (%).**

Dataset	Split	N	Acc@0.5	Acc@0.7	Mean IoU
RefCOCO	val	10,834	89.63	83.15	82.22
	testA	5,657	92.03	86.51	84.39
	testB	5,095	85.24	76.66	78.15
RefCOCO+	val	10,758	82.13	76.46	75.56
	testA	5,726	87.86	82.52	80.76
	testB	4,889	73.63	66.54	68.69
RefCOCOg	val	4,896	82.95	76.37	77.00
	test	9,602	83.63	77.57	77.99

G Data, Model Selection, and Evaluator Protocol

The official gRefCOCO manifests contain 209,344 training expressions, 14,229 validation expressions, 19,200 testA expressions, and 16,063 testB expressions. The controlled architecture study uses 12,000 training expressions and a disjoint 3,000-example development split with zero sample or image overlap with controlled train and official val/testA/testB.

3B model selection. The model used for final paper reporting is selected exclusively by official-validation Pr. Across the three training epochs, official-val Pr is 78.54, **81.24**, and 78.92%, respectively; the highest-validation model is used for testA/testB reporting. Neither test split participates in selection.

Table 15: **Validation-only model selection for the main 3B run.**

Training epoch	Official val Pr	Used for final reporting
1	78.54	–
2	**81.24**	**Yes**
3	78.92	–

7B scale protocol. The 7B scale study reports the model at completion of the pre-specified three-epoch training schedule (Pr 78.35, N-acc. 82.01). Test splits are not used for selection or tuning.

Evaluator integrity. The canonical paper evaluator retains malformed examples in the denominator and uses one-to-one IoU matching at threshold 0.5. Independent recomputation agrees with the training-time separate-decoder evaluator on all 3,000 development and 14,229 official-val predictions (0 disagreements out of 17,229). The final 3B summaries are recomputed from 49,492 cached official val/testA/testB predictions and reconcile with the paper tables to floating-point precision.

Pretrained-data statement. Qwen2.5-VL is a publicly pretrained vision-language model whose complete pretraining corpus is not exhaustively disclosed at the instance level. Task-specific adaptation in this work uses only the official gRefCOCO training annotations or the explicitly documented controlled subset. We therefore do not claim absence of generic COCO image exposure during backbone pretraining, while official testA/testB annotations are excluded from task-specific adaptation and model selection.

H Additional Qualitative Results



Figure 7: **Successful native complete-set grounding across cardinalities.** The examples span no-target, single-target, and increasingly large multi-target expressions.

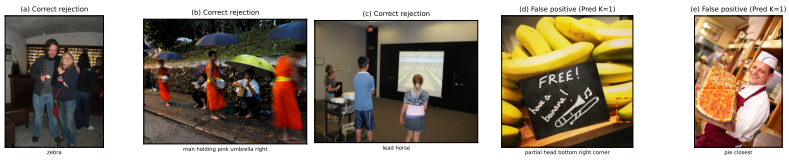

Figure 8: **Additional no-target behavior.** Correct rejections and false-positive no-target failures illustrate the behavior summarized by N-acc.