Subject-Agnostic Cross-View Referring Expression Comprehension via Reward-Driven Consistency Learning

Liuwu Li[1], **Ronger Ding**[1], **Zehang Lin**[3], **Qi Peng**[1] and **Jiayuan Xie**[2*]
[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
[3]School of Computer and Information Engineering Hanshan Normal University Chaozhou, China
jiayuan.xie@polyu.edu.hk

Abstract

Cross-view Referring Expression Comprehension (REC) requires a model to accurately localize objects in a spatial representation of one perspective based on descriptions generated from a different perspective. Essentially, it demands maintaining referential consistency under perspective shifts. This problem is prevalent in engineering application scenarios (e.g., Geographic Information Systems (GIS), urban planning, etc), where descriptions are typically generated from a top-down or planning view, while targets must be identified in a spatial representation distinct from the source. However, existing methods often assume shared perspectives or rely on an explicit observer to interpret observer-centric spatial relations (e.g., "left-/right" of the observer). Consequently, they struggle to handle cross-view grounding cases commonly arising in engineering application scenarios, where descriptions are defined by the source-view representation itself and do not involve any observing subject. To this end, we propose a novel Subject-Agnostic Cross-View REC task and construct a controlled simulation environment to systematically study referential consistency under perspective changes. Methodologically, we propose a Cross-View Bidirectional Reward Learning framework. By imposing mutually constraining reward signals on the grounding and localization results between the source and target views, we guide the model to maintain referential consistency across different view representations. Experimental results show that our method significantly outperforms baselines in multiple cross-view settings, exhibiting more stable generalization performance under perspective shifts.

1 Introduction

Referring expression comprehension (REC) aims to enable intelligent systems to uniquely identify the intended target object through natural language descriptions [Chen *et al.*, 2025]. As a fundamental task in multimodal perception and visual understanding, this task requires not only accurate semantic interpretation of language, but more importantly, the establishment of stable and reliable correspondences between linguistic descriptions and concrete objects. Such capabilities are essential for supporting perception, interaction, and decision-making in real-world applications. Existing studies have achieved significant progress on this task, largely under the assumption that natural language descriptions and perceptual inputs share the same view [Yang *et al.*, 2025; Deng *et al.*, 2021]. Under this same-view paradigm, the spatial relations and attributes expressed in referring expressions can be mapped to object features in the current perceptual view, enabling direct alignment between language and vision.

As intelligent systems are increasingly deployed in complex application scenarios, REC faces growing demands. Among these challenges, the inconsistency between the view in which referring expressions are generated and the perceptual view of target objects has become particularly prominent. To address this view inconsistency, a line of research has primarily focused on human–computer interaction scenarios, where an observing subject is explicitly modeled [Bu *et al.*, 2025]. By introducing the subject's position and orientation, referring expressions defined in the human view are reinterpreted as spatial relations in the machine view, with semantics relying on the subject-defined spatial reference frame, such as in expressions like "the second object to the left of the subject". However, beyond human–computer interaction, there exist a large number of subject-agnostic engineering application scenarios, including GIS, urban management, and engineering inspection. In these settings, referring expressions are not generated by any observing subject, but instead serve engineering purposes such as planning, localization, and inspection [Earl *et al.*, 2000; Gunes and Kovel, 2000; Wang *et al.*, 2025]. For example, in urban management systems, referring expressions can be defined on map-based or GIS-style views to indicate buildings that require inspection; in engineering inspection tasks, referring descriptions specified in top-down layouts can guide automated systems to locate specific facilities in ground-level perception views. As illustrated in Fig. 1, the referring expression "The curved building adjacent to the circular plaza on its left side, serving as the primary boundary structure of the plaza." is generated from a top-down map representation, whereas the target task is to lo-

*Corresponding author.

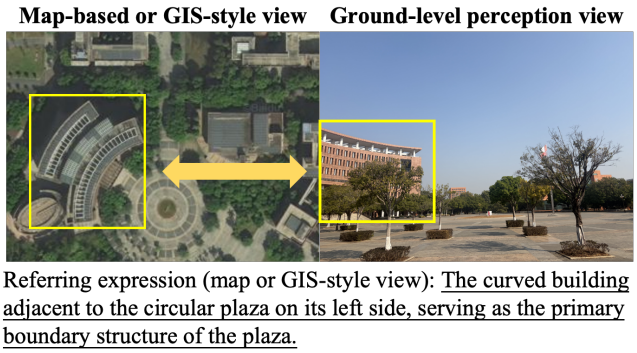


Figure 1: An example of subject-agnostic cross-view REC, where a target object is described in a map-based view and localized in a ground-level perception view.

calize the corresponding building in the ground-level perception view and perform subsequent search or maintenance operations. Under this setting, referential semantics are entirely defined by the geometric properties of objects and their relational structure with surrounding elements, rather than any subject-centered perspective. When the view transitions from a top-down representation to a ground-level perception view, referential semantics can no longer be grounded in a subject-based reference frame, but instead must be recovered through the alignment and reasoning over object relations across different view representations. As a result, existing subject-centric REC methods are difficult to apply directly.

In this paper, we focuse on subject-agnostic cross-view REC, which does not rely on any form of observer modeling. Unlike human–computer interaction scenarios, this setting emphasizes understanding and preserving referential semantic consistency across different view representations. A primary challenge of this task lies in the lack of large-scale datasets with consistent annotations in real-world scenarios. Application domains such as geographic information systems (GIS), urban management, and engineering inspection are often constrained by data acquisition costs, privacy and compliance requirements, as well as the complexity of synchronized multi-view collection, making it difficult to systematically construct paired multi-view images with corresponding referring expressions. To address this limitation, we construct a controlled simulated environment that generates diverse spatial configurations by composing modular elements with varying shapes, categories, and layouts. This construction follows a widely adopted practice in the community, as exemplified by synthetic datasets such as CLEVR [Johnson *et al.*, 2017]. Within this environment, we simultaneously collect top-down and frontal images, where referring expressions are generated based on the top-down view, while the target objects must be localized in the frontal view. It is important to emphasize that this simulated environment is not intended to simplify the task. Instead, it allows for systematic control over key structural factors involved in cross-view referring expression comprehension, e.g., object relations, global layout, and view variation, thereby avoiding confounding variables that are difficult to disentangle in real-world data.

Based on the above simulated environment, we propose a subject-agnostic cross-view referential grounding method built upon a multimodal large language model, termed ReCoG (Relational Consistency Grounding). Motivated by the strong performance of existing multimodal models under same-view grounding, we introduce language as an explicit intermediate representation to facilitate semantic alignment across different views. ReCoG adopts object-to-object relational consistency as its core modeling principle and leverages reinforcement learning to explicitly enforce the preservation and alignment of referential semantics across heterogeneous view representations, without relying on any form of observer modeling or explicit geometric transformations. Specifically, ReCoG first generates a referring description in the target view that depends solely on inter-object relations, thereby reconstructing the referential semantics defined in the source view. The model then localizes the target object based on this relation-consistent description. Through this decoupled formulation of relation reconstruction followed by object localization, ReCoG explicitly encourages the model to prioritize object relations that remain stable under viewpoint changes. To further ensure referential consistency across views, ReCoG introduces a reinforcement learning–based bidirectional consistency optimization mechanism. During training, the model is not only required to achieve accurate localization in the target view, but also to perform a reverse grounding process that mirrors the forward procedure: the relation-based referring description generated in the target view is fed back into the source view, where the model must ground it again to uniquely identify the original target object. By jointly enforcing reward constraints on both the forward and backward grounding processes, the model is compelled to learn referring expressions that are valid and consistent across different view representations, thereby achieving robust cross-view semantic alignment. Experimental results demonstrate that ReCoG maintains stable referential consistency under complex viewpoint changes and significantly outperforms direct cross-view grounding approaches.

To summarize, our contributions are as follows:

- We introduce, for the first time, a subject-agnostic cross-view referring expression comprehension problem, where referential semantics are defined in a planning-style top-down view while target objects must be localized in a ground-level perceptual view.
- We construct a structurally controlled cross-view referring expression dataset that covers object relations, global layouts, and view variations, providing a standardized benchmark for studying relation-driven cross-view grounding.
- We propose a relation-consistency–driven structured reinforcement learning approach that explicitly enforces bidirectional cross-view referential consistency, significantly outperforming same-view assumption methods under complex view transformations.

2 Related Work

REC aims to localize specific objects within visual content based on natural language descriptions [Qiao *et al.*, 2020; Li *et al.*, 2021; Wu *et al.*, 2023; Yuan *et al.*, 2025]. The

mainstream research paradigm predominantly operates under a same-view assumption, where the linguistic expression and the visual perception share an identical spatial reference frame [Yu *et al.*, 2016; Kamath *et al.*, 2021; Kang *et al.*, 2025; Wang *et al.*, 2024]. While effective for standard retrieval, these methods struggle in realistic engineering and embodied scenarios where the viewpoint of description generation deviates from that of visual perception [Li *et al.*, 2016; Pramanick *et al.*, 2022; Yuan *et al.*, 2026b]. In such cases, the spatial relations encoded in language cannot be directly mapped onto the current perceptual view, creating a significant gap in semantic alignment.

To address viewpoint disparities, recent studies have expanded into cross-view settings, often focusing on reconciling disparate perspectives between a speaker and an agent [Pramanick *et al.*, 2022; Chen *et al.*, 2023; Yuan *et al.*, 2026a]. For example, [Huang *et al.*, 2022] propose language-driven robot manipulation that resolves perspective ambiguity through scene descriptions. More recently, [Bu *et al.*, 2025] introduce a human-centric REC task, employing a Top-view Enhanced Perspective transformation to interpret instructions generated from a human's perspective that differs from the robot's egocentric view. However, these approaches are fundamentally observer-centric. They rely on modeling a specific subject to interpret relative directional commands (e.g., "on my left"), limiting their applicability in subject-agnostic domains like GIS or engineering inspection, where descriptions are view-defined rather than user-defined. Distinguishing our work from observer-centric methods, we address a subject-agnostic cross-view REC task where referential semantics rely solely on structural object-to-object relations defined in the source view, independent of any observer's pose.

3 Data Construction

In this paper, we construct a controlled cross-view referential dataset. The dataset is designed to simulate the discrepancy between map-based or GIS-style representations and ground-level perception views, providing a controllable and scalable testbed for analyzing cross-view grounding mechanisms.

Cross-view Scene and Referring Expression Construction. Ideally, subject-agnostic cross-view referential understanding should be investigated using real-world data from domains such as GIS and engineering inspection. However, constructing large-scale datasets for such settings remains impractical due to the high cost of synchronized multi-view acquisition, strict privacy and compliance constraints, and the lack of consistent cross-view referential annotations. Given these limitations, we construct cross-view referential scenes in a controlled laboratory environment. Each scene is composed of blocks with diverse geometric shapes, colors, and sizes, arranged in non-regular layouts to form varied spatial configurations. This design avoids rigid templates and introduces object interference and relational ambiguity observed in real-world engineering scenarios.

For each scene, two views are captured simultaneously: i) a *source (top-down) view*, simulating map-based, planning, GIS-style, or schematic representations, and ii) a *target (frontal) view*, simulating ground-level perception observations. No explicit geometric alignment or coordinate transformation between the two views is provided, preventing models from relying on simple viewpoint mappings.

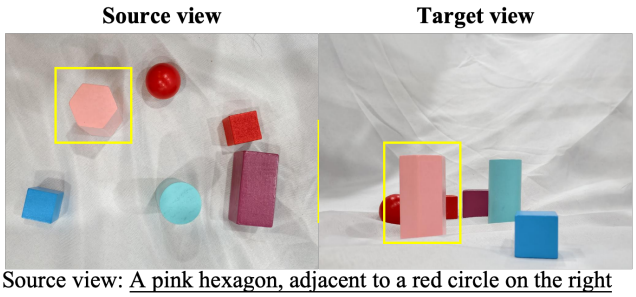


Figure 2: A sample of a dataset we constructed.

Referring expressions are generated exclusively based on the top-down view. For each scene, a target object is first selected in the source view, followed by annotating a natural-language expression that uniquely identifies the target. Importantly, all expressions strictly follow a subject-agnostic principle and do not include any observing subject, orientation information, or first-person references. Instead, referential semantics are anchored to object-to-object relations defined in the source-view layout.

Each referring expression contains one or more of the following information dimensions: i) appearance attributes of the target object (e.g., shape, color, or relative size); ii) the target object's relative region or position in the top-down layout; iii) relational constraints between the target object and other distinguishable objects (e.g., adjacency or relative placement). This protocol ensures that referential semantics encode structural relations rather than subjective viewpoints.

Dataset Scale and Structural Difficulty Split The constructed dataset contains a total of 5,900 samples. Each sample consists of a source view image, a target view image, and one referring expression. Figure 2 illustrates a representative example. The referring expression is generated from the top-down view based on object geometry and spatial relations, while the target object must be localized in the frontal view, where appearance and visibility differ substantially. This example highlights the core challenge of subject-agnostic cross-view referential grounding studied in this work.

To systematically evaluate model performance under different levels of structural complexity, we divide the dataset based on the number of objects present in each scene: i) *sparse scenes*, where the number of objects is no greater than four, and ii) *dense scenes*, where more than four objects are present, resulting in increased relational complexity and object interference. Under this split, the dataset is composed of: 612 sparse scenes, all of which are used for testing; and 5,288 dense scenes, from which 612 scenes are randomly selected for testing and others are used for training.

This setting ensures that the test set covers both low- and high-complexity scenarios, while training primarily relies on structurally complex scenes, enabling a more rigorous evaluation of cross-view relational generalization.

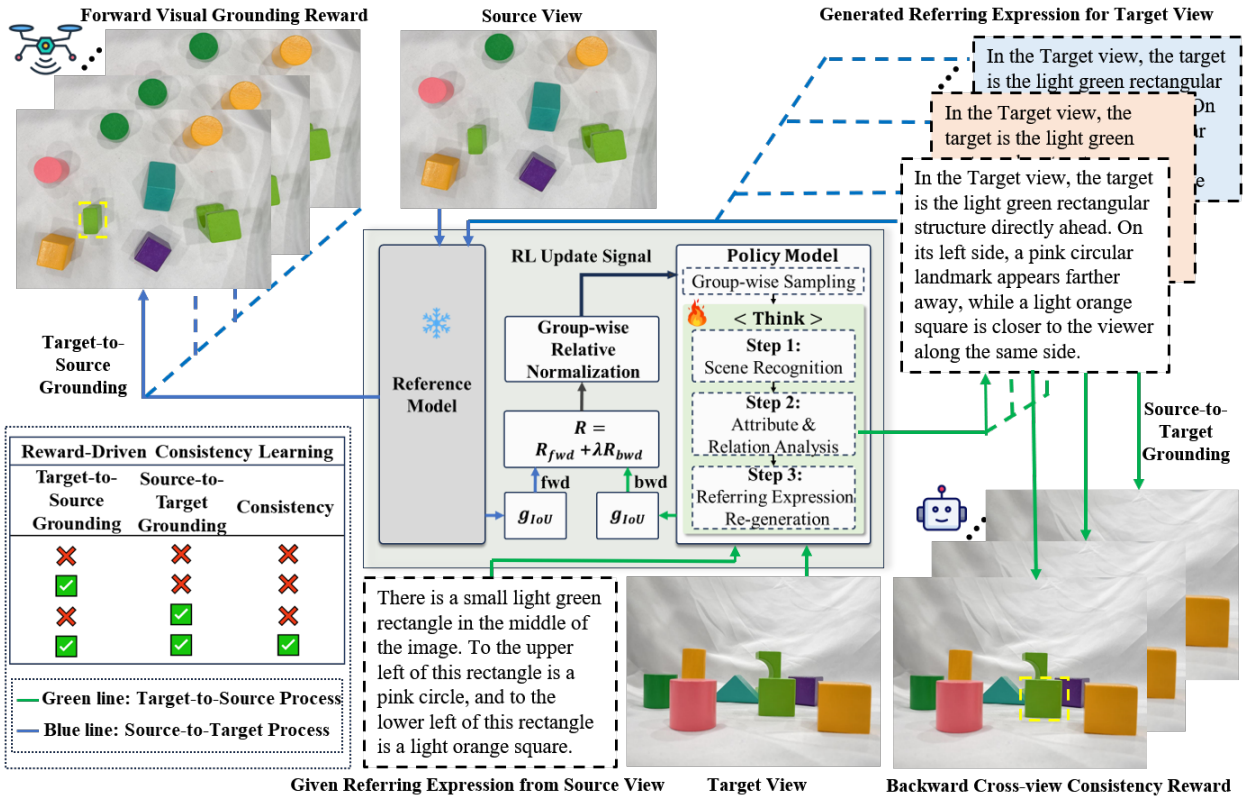


Figure 3: Overview of ReCoG. It performs subject-agnostic cross-view REC via a bidirectional generation-and-grounding pipeline.

4 Method

In this paper, we aim to align linguistic semantics with target objects under mismatched viewpoints between expression generation and visual perception. Formally, given a referring expression q_s generated from a source-view representation (e.g., a top-down view, a planning view), and a visual observation I_t captured from a target perception view (e.g., a front-facing or local view), the goal is to localize the referred object in I_t and output its corresponding spatial representation b_t, such as a bounding box or coordinate position.

4.1 Relational Consistency Grounding (ReCoG)

We introduce Relational Consistency Grounding (ReCoG), a subject-agnostic framework for cross-view referential grounding. As illustrated in Figure 3, the core principle of ReCoG is to explicitly enforce the preservation of relational semantics across heterogeneous viewpoints through a bidirectional consistency mechanism.

Specifically, ReCoG decomposes cross-view grounding into a forward–backward semantic loop. In the forward direction (green arrows in Figure 3), the model transfers the referring semantics defined in the source view to the target view, reconstructs the relational representation of the target object under the new viewpoint, and predicts its spatial localization accordingly. In the backward direction (blue arrows), the model maps the target-view relational description back to the source view and verifies whether it can still uniquely re-identify the original object. This closed-loop formulation forces the model to learn object-centric relational representations that remain stable under viewpoint changes, thereby enabling genuine cross-view semantic alignment rather than view-specific pattern matching.

Formally, given a source-view referring expression q_s and a target-view image I_t, the objective is to localize the corresponding object in the target view. Instead of producing a single deterministic prediction, we cast cross-view grounding as a stochastic decision-making process. A model parameterized by θ defines a conditional policy

$$\pi_\theta(b_t \mid q_s, I_t), \tag{1}$$

where b_t denotes a candidate target-view localization. During training, multiple localization hypotheses are sampled under the same input condition, allowing the model to explore diverse cross-view alignment possibilities (see Figure 3).

Crucially, for each sampled localization, the model additionally generates an intermediate target-view referring description that characterizes the object in terms of its attributes and relations to surrounding entities. This description is strictly subject-agnostic, avoiding observer-dependent spatial references, and serves as a compact semantic abstraction of the object under the target viewpoint. Notably, this intermediate description is not supervised explicitly as a reasoning trace; instead, it functions as a latent semantic carrier through which cross-view consistency is enforced.

To operationalize relational consistency, ReCoG intro-

duces a reinforcement learning-based bidirectional grounding objective. During training, the target-view intermediate description generated from a candidate localization is grounded back to source view, where the model attempts to re-localize the original object. This forward (source-to-target) and backward (target-to-source) grounding procedure forms a closed semantic loop. Candidate predictions are rewarded only if they yield consistent object localization across both views, thereby discouraging spurious or view-specific alignments.

The overall framework is optimized using structured reinforcement learning with group-wise sampling, where multiple candidate predictions are evaluated relative to one another under identical input conditions. This design enables stable and efficient learning of cross-view relational grounding without requiring explicit cross-view correspondence supervision. At inference time, ReCoG performs only the forward grounding process. Following the structured reasoning pathway illustrated in the framework (i.e., the ⟨Think⟩ block), the model generates a relation-driven target-view referring description and predicts the object localization accordingly, without invoking backward verification.

4.2 Reward Design

To optimize subject-agnostic cross-view REC under limited supervision, we design reward functions that explicitly enforce cross-view localization consistency.

Forward Visual Grounding Reward Given an input pair (q_s, I_t), our policy model generates a set of candidate target-view predictions via group-wise sampling. Although the policy model itself is capable of producing end-to-end localization outputs through a thought chain process (i.e., the <Think> block), directly evaluating its own predictions during reinforcement learning would tightly couple generation and reward signals, leading to unstable optimization. To provide a stable and reliable evaluation signal, we introduce a fixed reference model, which is a strong multimodal model trained for same-view referring expression grounding.

Specifically, the expression generated by the policy model is passed to the reference model together with the target-view image I_t, and the reference model performs same-view grounding to produce a predicted bounding box $b_t^{(k)}$. The forward grounding reward is then computed by comparing this prediction with the ground-truth target-view localization b_t^* using generalized Intersection-over-Union (gIoU):

$$R_{\text{fwd}}^{(k)} = \text{gIoU}\big(b_t^{(k)},\, b_t^*\big). \tag{2}$$

Compared to standard IoU, gIoU provides informative gradients even when predicted and ground-truth boxes do not overlap, thereby stabilizing reinforcement learning optimization. Importantly, this reward evaluates the quality of cross-view relation reconstruction, rather than replacing the localization capability of the policy model itself.

Backward Cross-view Consistency Reward To further enforce cross-view semantic consistency, we introduce a backward consistency reward that mirrors the forward grounding process. For each candidate prediction, the policy model generates a target-view referring expression as described above. This expression is then grounded back to the source view by the Reference Model, producing a source-view localization $\hat{b}_s^{(k)}$.

The backward reward evaluates whether this reverse grounding recovers the source-view target localization b_s^*:

$$R_{\text{bwd}}^{(k)} = \text{gIoU}\big(\hat{b}_s^{(k)},\, b_s^*\big). \tag{3}$$

The forward and backward grounding processes are structurally symmetric. A candidate prediction receives a high backward reward only if the generated target-view description remains valid when mapped back to the source view, thereby uniquely identifying the original target object.

The policy model remains responsible for cross-view semantic modeling and final target localization, while the reference model is used exclusively during training as a frozen evaluator that provides stable same-view grounding feedback. The IoU-based rewards obtained from the reference model reflect how well the policy model's generated referring expressions encode cross-view relational semantics.

Overall Reward The overall reward for each candidate localization is defined as a weighted combination of the forward grounding reward and the backward consistency reward:

$$R^{(k)} = R_{\text{fwd}}^{(k)} + \lambda\, R_{\text{bwd}}^{(k)}, \tag{4}$$

where λ controls the contribution of backward consistency. In practice, λ is set to a moderate value to ensure that target-view localization accuracy remains the dominant objective while still enforcing cross-view consistency.

Group-wise Relative Optimization For the optimization of RL, we adopt a group-wise relative optimization strategy to reduce sensitivity to reward scale. For a candidate set $\mathcal{B}$ generated under the same input condition, we compute the group-wise mean and standard deviation of rewards:

$$\mu = \frac{1}{K}\sum_{k=1}^{K} R^{(k)}, \qquad \sigma = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\big(R^{(k)} - \mu\big)^2}. \tag{5}$$

The normalized relative advantage for the $b_t^{(k)}$ is defined as:

$$A^{(k)} = \frac{R^{(k)} - \mu}{\sigma + \epsilon}, \tag{6}$$

where ϵ is a small constant for numerical stability. Policy updates are guided by these relative advantages, encouraging candidates that outperform others within the same group and improving training stability and sample efficiency.

4.3 Training and Inference Procedure

During training, the model is optimized with structured reinforcement learning using group-wise sampling. For each sample, multiple target-view localization candidates are generated under the same input and jointly evaluated by the reward functions in Section 4.2, which assess grounding quality and cross-view consistency through forward–backward grounding. Model parameters are updated with group-wise relative advantages, while curriculum learning progressively introduces more structurally complex scenes.

During inference, the policy model first performs chain-of-thought reasoning, then predicts the target object localization in the target view given a source-view referring expression and a target-view image.

Model	Scene Type	mIoU	Acc@0.3	Acc@0.5	Acc@0.7	Acc@0.9
Claude Haiku 4.5 [Anthropic, 2025]	Sparse	17.84	25.82%	24.18%	11.76%	0.00%
	Dense	26.18	40.52%	29.74%	14.05%	0.33%
GPT5_mini [OpenAI, 2025]	Sparse	29.33	51.10%	20.96%	4.04%	0.00%
	Dense	31.74	53.75%	25.30%	4.35%	0.40%
Gemini-3-Flash [Gemini, 2025]	Sparse	21.76	29.73%	28.38%	17.57%	0.00%
	Dense	19.67	25.42%	22.71%	17.63%	0.34%
Qwen3-VL-2B [Bai *et al.*, 2025]	Sparse	33.51	43.14%	43.14%	35.29%	0.00%
	Dense	38.91	47.06%	47.06%	43.79%	4.58%
Qwen3-VL-4B [Bai *et al.*, 2025]	Sparse	22.25	28.76%	28.76%	24.84%	0.00%
	Dense	29.34	36.27%	35.95%	33.99%	2.61%
Qwen3-VL-8B [Bai *et al.*, 2025]	Sparse	44.44	58.17%	58.17%	49.02%	0.00%
	Dense	41.49	51.63%	50.98%	47.39%	3.59%
Qwen3-VL-2B-ft	Sparse	65.03	73.53%	73.53%	73.20%	28.07%
	Dense	66.79	74.18%	73.86%	73.86%	30.18%
Qwen3-VL-2B-RL	Sparse	67.02	76.47%	76.47%	76.14%	29.32%
	Dense	68.45	76.14%	75.49%	75.49%	27.82%
ReCoG-2B w/o Generated RefExp	Sparse	66.92	76.88%	75.44%	73.02%	22.14%
	Dense	68.94	78.02%	76.88%	75.44%	31.26%
ReCoG-2B w/o Backward Consistency Reward	Sparse	63.84	73.12%	71.94%	69.21%	18.37%
	Dense	65.27	74.83%	73.41%	70.92%	25.11%
ReCoG-2B	Sparse	<u>71.03</u>	<u>80.56%</u>	<u>80.56%</u>	<u>79.90%</u>	**32.52%**
	Dense	<u>73.38</u>	<u>82.19%</u>	<u>81.70%</u>	<u>81.21%</u>	**43.46%**
ReCoG-4B	Sparse	**80.92**	**92.81%**	**92.81%**	**92.16%**	<u>29.52%</u>
	Dense	**76.69**	**86.60%**	**86.60%**	**85.78%**	<u>37.75%</u>

Table 1: Model performance comparison on sparse and dense scenes. For each scene type (Sparse/Dense), the best result in each column is highlighted in bold, and the second-best result is <u>underlined</u>.

5 Experiments

Setting. We adopt Qwen3-VL [Bai *et al.*, 2025] as the backbone of our ReCoG. During the reinforcement learning phase, sixteen candidate outputs are generated for each sample for relative reward calculation and policy optimization, thereby encouraging the model to explore a variety of cross-view alignment strategies.

Evaluation Metrics. Following research in the REC field [Li *et al.*, 2025], we use mIoU and Acc@τ at different thresholds to evaluate model performance. The mIoU measures the average overlap between predicted bounding boxes and ground-truth boxes across the test set. Acc@τ denotes the proportion of samples whose predicted IoU exceeds a given threshold τ (e.g., 0.3, 0.5, 0.7, and 0.9).

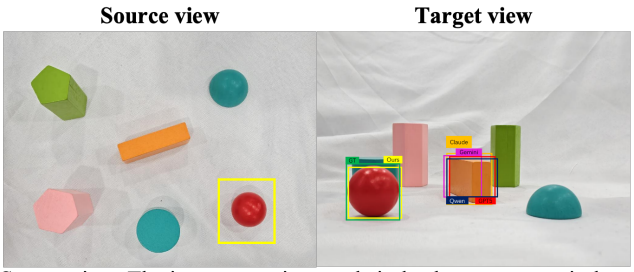
Comparison Models and Variants. We compare our method with several categories of models, including Pretrained multimodal models, a fine-tuned model, and reinforcement learning-based variants.

Pretrained multimodal models, including Claude Haiku 4.5 [Anthropic, 2025], GPT-5-mini [OpenAI, 2025], Gemini-3-Flash [Gemini, 2025], and Qwen3-VL-2B/4B/8B are evaluated as zero-shot baselines to assess the generic cross-view grounding capability of current systems without task-specific adaptation. Qwen3-VL-2B-fine-tuned models (ft) further adapt the backbone to the proposed task using supervised learning, but still rely on direct grounding without enforcing cross-view relational consistency. The Qwen3-VL-2B-RL adopts GRPO-based [Shao *et al.*, 2024] single-direction reinforcement learning [Yan *et al.*, 2025], where target-view localization is optimized conditioned on source-view REC only. In contrast, our ReCoG-RL introduces structured relational reasoning and bidirectional consistency training.

To further analyze the contribution of each component, we conduct two ablation studies. ReCoG-2B w/o Generated RefExp removes the intermediate generation of target-view referring expressions and directly predicts bounding boxes, disabling explicit relational grounding through language. ReCoG-2B w/o Backward Consistency Reward removes the backward grounding process from the target view back to the source view, eliminating bidirectional consistency constraints during reinforcement learning.

5.1 Results and Analysis

Table 1 summarizes the quantitative results of different models on both Sparse and Dense scenarios. We will analyze the results from the following aspects.



Source view: The image contains a red circle above a green circle, with an orange rectangle at the bottom left corner of the green circle.

Figure 4: A sample of a dataset we constructed.

Sparse vs. Dense Scenes. Dense scenes pose substantially greater challenges than Sparse scenes due to increased object interference and relational ambiguity. Most baseline models exhibit notable performance degradation under Dense settings, particularly at stricter localization thresholds. For example, Claude Haiku 4.5 achieves only 0.33% Acc@0.9 in Dense scenes, while GPT5_mini and Gemini-3-Flash show similarly limited high-IoU performance. In contrast, our ReCoG models maintain strong robustness under Dense conditions. ReCoG-2B achieves 73.38 mIoU and 43.46% Acc@0.9 in Dense scenes, demonstrating that relation-driven consistency modeling is especially effective in structurally complex environments.

General-Purpose Pre-trained Multimodal Models. General-purpose multimodal large models perform poorly in zero-shot cross-view grounding. Although GPT5_mini and Gemini-3-Flash attain moderate Acc@0.3 scores, their Acc@0.7 and Acc@0.9 rapidly collapse (e.g., below 5% in Sparse scenes for GPT5_mini). These results indicate that generic vision–language capabilities alone are insufficient for reliable cross-view referential localization without explicit relational alignment mechanisms.

Effect of Fine-Tuning. Task-specific fine-tuning substantially improves performance over pretrained models. Qwen3-VL-2B-ft improves mIoU from 33.51 to 65.03 in Sparse scenes and reaches 66.79 in Dense scenes. However, its high-threshold localization remains limited (Dense Acc@0.9 at 30.18%), suggesting that supervised fine-tuning alone cannot fully enforce cross-view relational consistency.

Model Parameter Scale Comparison. Increasing model parameter scale does not guarantee better cross-view grounding. While Qwen3-VL-8B improves over smaller pretrained variants in Sparse scenes (44.44 mIoU), its Dense-scene performance degrades to 41.49 mIoU with Acc@0.9 remaining below 4%. In contrast, ReCoG-2B and ReCoG-4B consistently outperform larger pretrained models, highlighting that relational consistency modeling is more critical than parameter scale for this task.

Comparison of Reinforcement Learning Strategies. Different reinforcement learning designs lead to markedly different outcomes. Qwen3-VL-2B-RL employs single-direction GRPO optimization and achieves 27.82% Acc@0.9 in Dense scenes. ReCoG further introduces structured reasoning and bidirectional consistency training, explicitly enforcing referential alignment across views. This results in substantial gains, with ReCoG-2B improving Dense Acc@0.9 to 43.46% and ReCoG-4B reaching 92.16% Acc@0.7 in Sparse scenes. These results confirm that reward-driven relational consistency is the key factor behind the performance improvements.

Ablation Analysis We further analyze the contribution of key components in ReCoG through two ablation settings. Removing the intermediate generation of target-view referring expressions (ReCoG-2B w/o Generated RefExp) leads to a clear performance drop, with mIoU decreasing from 71.03 to 66.92 on sparse scenes and from 73.38 to 68.94 on dense scenes. This indicates that explicitly reconstructing relation-consistent textual descriptions plays a critical role in bridging cross-view semantic gaps. When the backward consistency reward is removed (ReCoG-2B w/o Backward Consistency Reward), performance further degrades (mIoU 63.84 / 65.27 on sparse/dense scenes), demonstrating that enforcing bidirectional grounding consistency is essential for maintaining stable referential semantics across views.

6 Case Study

Figure 4 illustrates a representative cross-view referring example comparing our method with several pretrained baselines. As shown in the right panel, pretrained multimodal models—including GPT-5, Claude, Gemini, and Qwen—fail to localize the correct target and instead drift toward visually salient but relationally inconsistent objects. This behavior stems from their reliance on local appearance cues and weak spatial heuristics, without an explicit mechanism to enforce relational consistency across views, leading to ambiguous or incorrect grounding under viewpoint changes.

In contrast, our method accurately localizes the red spherical object by explicitly modeling object-to-object relational structure rather than relying on isolated appearance cues or keyword matching in the referring expression. Crucially, the effectiveness of our approach stems from its bidirectional consistency mechanism: the model is required not only to ground a relation-consistent description in the target view, but also to ensure that this description can be grounded back to the source view and still uniquely identify the original target.

7 Conclusion

This paper addresses the fundamental challenge of preserving relational semantics under viewpoint changes in subject-agnostic cross-view REC. We propose a structured reinforcement learning framework that explicitly models relation-driven cross-view semantics and enforces bidirectional consistency, enabling referential meanings to remain valid when transferred between heterogeneous view representations. To support systematic analysis, we further construct a controlled cross-view referential dataset with varying structural complexity. Extensive experiments demonstrate that our approach consistently outperforms pretrained multimodal models and existing reinforcement learning variants, especially in structurally complex scenes, highlighting the effectiveness of relation-consistency–driven and subject-agnostic modeling for cross-view referential understanding.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (62476097), the Fundamental Research Funds for the Central Universities, South China University of Technology (x2rjD2250190), Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078), Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010).

References

[Anthropic, 2025] Anthropic. Claude haiku 4.5 system card. https://assets.anthropic.com/m/99128ddd009bdcb/claude-haiku-4-5-system-card.pdf. 2025.

[Bai *et al.*, 2025] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[Bu *et al.*, 2025] Yuqi Bu, Xin Wu, Zirui Zhao, Yi Cai, David Hsu, and Qiong Liu. Walk in others' shoes with a single glance: Human-centric visual grounding with top-view perspective transformation. In *Proc. of ACL*, pages 26904–26923, 2025.

[Chen *et al.*, 2023] Kaiqi Chen, Jing Yu Lim, Kingsley Kuan, and Harold Soh. Latent emission-augmented perspective-taking (leapt) for human-robot interaction. In *2023 IROS*, pages 8006–8013. IEEE, 2023.

[Chen *et al.*, 2025] Xiaofu Chen, Yaxin Luo, Gen Luo, Jiayi Ji, Henghui Ding, and Yiyi Zhou. Dvin: Dynamic visual routing network for weakly supervised referring expression comprehension. In *Proc. of CVPR*, pages 14347–14357. Computer Vision Foundation / IEEE, 2025.

[Deng *et al.*, 2021] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *In Proc. of ICCV*, pages 1769–1779, 2021.

[Earl *et al.*, 2000] R Earl, G Thomas, and BS Blackmore. The potential role of gis in autonomous field operations. *Computers and electronics in agriculture*, 25(1-2):107–120, 2000.

[Gemini, 2025] Gemini. Gemini-3-flash. gemini-3-flash. https://blog.google/products/gemini/gemini-3-flash/. 2025.

[Gunes and Kovel, 2000] A Ertug Gunes and Jacob P Kovel. Using gis in emergency management operations. *Journal of Urban Planning and Development*, 126(3):136–149, 2000.

[Huang *et al.*, 2022] Kaixiang Huang, Yuxuan Han, Jinze Wu, Fangzhou Qiu, and Qing Tang. Language-driven robot manipulation with perspective disambiguation and placement optimization. *IEEE Robotics and Automation Letters*, 7(2):4188–4195, 2022.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. of CVPR*, pages 2901–2910, 2017.

[Kamath *et al.*, 2021] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multimodal understanding. In *Proc. of ICCV*, pages 1780–1790, 2021.

[Kang *et al.*, 2025] 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 *In Proc. of CVPR*, pages 9339–9350, 2025.

[Li *et al.*, 2016] Shen Li, Rosario Scalise, Henny Admoni, Stephanie Rosenthal, and Siddhartha S Srinivasa. Spatial references and perspective in natural language instructions for collaborative manipulation. In *Proc. of 2016 25th IEEE RO-MAN*, pages 44–51. IEEE, 2016.

[Li *et al.*, 2021] Liuwu Li, Yuqi Bu, and Yi Cai. Bottom-up and bidirectional alignment for referring expression comprehension. In *Proc. of the 29th ACM MM*, pages 5167–5175, 2021.

[Li *et al.*, 2025] Liuwu Li, Zhuoming Zheng, Yuqi Bu, Cantao Wu, Shubin Huang, Qingbao Huang, and Yi Cai. Grouped top-down reasoning with hierarchical window transformer for visual grounding. *Information Processing & Management*, 62(6):104222, 2025.

[OpenAI, 2025] OpenAI. Gpt-5 mini – openai api model documentation. https://platform.openai.com/docs/models/gpt-5-mini/. 2025.

[Pramanick *et al.*, 2022] Pradip Pramanick, Chayan Sarkar, Sayan Paul, Ruddra dev Roychoudhury, and Brojeshwar Bhowmick. Doro: Disambiguation of referred object for embodied agents. *IEEE Robotics and Automation Letters*, 7(4):10826–10833, 2022.

[Qiao *et al.*, 2020] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020.

[Shao *et al.*, 2024] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[Wang *et al.*, 2024] Suixue Wang, Zhigao Zheng, Xinghong Wang, Qingchen Zhang, and Zhuo Liu. A cloud-edge collaboration framework for cancer survival prediction to develop medical consumer electronic devices. *IEEE Transactions on Consumer Electronics*, 70(3):5251–5258, 2024.

[Wang *et al.*, 2025] Suixue Wang, Shilin Zhang, Huiyuan Lai, Weiliang Huo, and Qingchen Zhang. Pomp: Pathology-omics multimodal pre-training framework for cancer survival prediction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7813–7821, 2025.

[Wu *et al.*, 2023] Cantao Wu, Yi Cai, Liuwu Li, and Jiexin Wang. Scene graph enhanced pseudo-labeling for referring expression comprehension. In *Findings of the EMNLP*, pages 11978–11990, 2023.

[Yan *et al.*, 2025] Qingyang Yan, Guangyao Chen, and Yixiong Zou. Start small, think big: Curriculum-based relative policy optimization for visual grounding. *arXiv preprint arXiv:2511.13924*, 2025.

[Yang *et al.*, 2025] Xuzheng Yang, Junzhuo Liu, Peng Wang, Guoqing Wang, Yang Yang, and Heng Tao Shen. New dataset and methods for fine-grained compositional referring expression comprehension via specialist-mllm collaboration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(10):8598–8612, 2025.

[Yu *et al.*, 2016] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proc. of ECCV 2016*, pages 69–85. Springer, 2016.

[Yuan *et al.*, 2025] Li Yuan, Yi Cai, Xudong Shen, Qing Li, Qingbao Huang, Zikun Deng, and Tao Wang. Collaborative multi-lora experts with achievement-based multi-tasks loss for unified multimodal information extraction. In *International Joint Conference on Artificial Intelligence*, 2025.

[Yuan *et al.*, 2026a] Li Yuan, Yi Cai, Bingshan Zhu, Zhenghao Liu, Zikun Deng, Qing Li, and Tao Wang. Visual knowledge-enhanced llava for fine-grained multimodal named entity recognition and grounding. *IEEE Transactions on Audio, Speech and Language Processing*, 34:781–795, 2026.

[Yuan *et al.*, 2026b] Li Yuan, Qingfei Huang, Bingshan Zhu, Yi Cai, Qingbao Huang, Changmeng Zheng, Zikun Deng, and Tao Wang. Hybrid-dmkg: A hybrid reasoning framework over dynamic multimodal knowledge graphs for multimodal multihop qa with knowledge editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28032–28040, 2026.