CityVG: Contrastive Fine-Tuning and Reward-Based Chain-of-Thought Reasoning for Zero-Shot City-Scale 3D Visual Grounding

Jianjun Zhang[1], **Hanli Wang**[1,2,3]*
[1]School of Computer Science and Technology, Tongji University, Shanghai, China
[2]College of Electronic and Information Engineering, Tongji University, Shanghai, China
[3]Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China
jianjunzhang@tongji.edu.cn, hanliwang@tongji.edu.cn

Abstract

3D Visual Grounding (3DVG) locates objects in 3D scenes based on natural language descriptions. However, existing methods are primarily confined to small-scale indoor data or rely on heavy supervision, failing to generalize to complex large-scale urban environments. To address this limitation, we present CityVG, the first city-scale zero-shot 3D visual grounding framework capable of localizing urban objects without manual annotations. Our approach adopts a retrieval-and-reasoning paradigm comprising two key components. Specifically, we propose a contrastive fine-tuning strategy to align textual queries with urban scene graphs. By leveraging an LLM-driven graph clustering mechanism, we automatically construct high-quality positive and negative training pairs and fine-tune the text encoder via contrastive learning, resulting in a scene-adaptive text encoder that enables efficient alignment without grounding supervision. Complementing this, a multi-trajectory reward-based Chain-of-Thought (CoT) reasoning strategy is designed for inference. This mechanism iteratively evaluates candidate objects by aggregating reward scores across diverse reasoning trajectories, selecting the target that is most consistent with both appearance and spatial constraints. Extensive experiments on city-scale 3D grounding benchmarks demonstrate that CityVG achieves strong zero-shot localization performance and generalizes effectively to unseen urban environments. The source code of this work can be found in https://mic.tongji.edu.cn.

1 Introduction

3D Visual Grounding locates objects (Zhang et al., 2024; Liu et al., 2026) in 3D scenes based on natural language descriptions, serving as a critical bridge for robotics (Xiao et al., 2026; Garcia et al., 2025), augmented reality (Hourcade et al., 2024), and autonomous navigation (Kong et al., 2025; Zhang et al., 2025). Extending this capability from room-scale indoor scenes to city-scale point clouds unlocks detailed urban analysis, allowing systems to interpret complex city layouts for urban planning and geo-spatial intelligence.

However, the transition to city-scale environments imposes severe limitations on current methodologies. The most immediate barrier is data scarcity: acquiring annotated point clouds for vast urban areas is prohibitively expensive, making manual labeling practically infeasible. Beyond data constraints, the scale itself is a challenge. Unlike indoor scenes (Armeni et al., 2016; Dai et al., 2017) with limited objects, urban environments (Hu et al., 2022; Lin et al., 2022) contain thousands of entities, rendering the exhaustive object–query matching strategies used in traditional methods computationally intractable. This is further complicated by intricate spatial semantics, where hierarchical and long-range relationships defy the local reasoning capabilities of existing parsers.

Existing methods struggle to address these scale-specific hurdles. Although vision-language models (VLMs) have improved multi-modal alignment, they are predominantly tailored for small-scale indoor settings (Yuan et al., 2024; Xu et al., 2025a; Li et al., 2025b). Recent attempts to bridge this gap, such as CityRefer (Miyanishi et al., 2023) and CityAnchor (Li et al., 2025a), employ supervised multi-modal embeddings and coarse-to-fine pipelines. Crucially, these approaches remain tied to supervised training. Consequently, they inherit the high annotation costs and computational bottlenecks mentioned above, limiting their ability to scale efficiently to zero-shot scenarios.

To overcome these limitations, we present CityVG, the first zero-shot framework for city-scale 3D visual grounding (Fig. 1). We address the dual challenges of scalability and supervision through a retrieval-and-reasoning paradigm. In the retrieval stage, we employ a contrastive fine-tuning strat-

*Corresponding author: Hanli Wang.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 35012–35031

egy anchored by a Visual Pairwise Scene Graph. By leveraging LLM-driven graph clustering to automatically generate pseudo-supervision, we efficiently align textual queries with urban structures, significantly pruning the search space without manual annotations. In the reasoning stage, we mitigate ambiguity via a multi-trajectory reward-based Chain-of-Thought strategy. By aggregating reward scores across diverse reasoning paths, the model suppresses spurious matches and robustly identifies targets that satisfy both appearance and spatial constraints within complex city layouts.

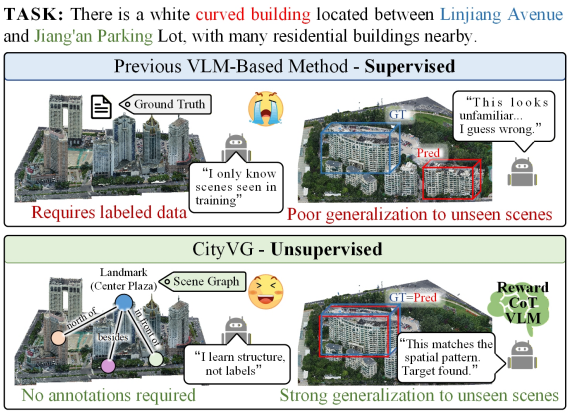

Figure 1: Illustration of previous VLM-based methods and CityVG. Compared with previous methods, CityVG performs grounding without annotations and demonstrates strong generalization to unseen scenes.

Our contributions are summarized as follows:

- We propose CityVG, the first zero-shot framework for city-scale 3D visual grounding. It introduces a retrieval-and-reasoning paradigm to localize in massive urban point clouds without requiring manual annotations.
- We design a contrastive fine-tuning strategy driven by an LLM-based graph clustering mechanism. This approach automatically generates training pairs to align textual queries with urban scenes, efficiently pruning the vast search space.
- We introduce a multi-trajectory reward-based Chain-of-Thought strategy. By evaluating candidates across diverse reasoning paths, we mitigate ambiguity and ensure robust localization under complex spatial constraints.
- CityVG achieves superior zero-shot performance on city-scale benchmarks, effectively generalizing to unseen environments and overcoming the scalability limitations of existing supervised methods.

2 Related Work

2.1 3D Visual Grounding

Research on 3D visual grounding has evolved from supervised feature alignment to advanced reasoning-based paradigms. Early approaches (Yuan et al., 2021; Zhao et al., 2021; Wu et al., 2023; Qian et al., 2024; Unal et al., 2024) typically map point cloud features and textual descriptions into a shared embedding space, enabling grounding via similarity matching. Subsequent works (Yang et al., 2024; Yuan et al., 2024; Xu et al., 2025a; Li et al., 2025b) have integrated VLMs and LLMs to enhance semantic understanding for free-form queries. While these innovations improve flexibility, they are predominantly tailored to indoor environments with limited object counts. Recently, the field has begun to address the complexities of city-scale grounding. Notable efforts such as CityRefer (Miyanishi et al., 2023) and CityAnchor (Li et al., 2025a) extend the task to large urban point clouds, utilizing supervised multi-modal embeddings and coarse-to-fine pipelines, respectively. Despite these advances, such methods remain heavily reliant on supervision and complex reasoning architectures, which constrain their ability to scale efficiently or generalize to zero-shot scenarios in unseen urban environments.

2.2 Language-Driven and Graph-Based Representation Learning

Graph-based representations have been widely adopted for modeling structured semantics in 3D scenes. Pioneering works like 3D Scene Graphs (Armeni et al., 2019; Kim et al., 2019) encode objects as nodes and relations as edges, a structure subsequently adopted by vision–language frameworks such as VL-SAT (Wang et al., 2023b) and Open3DSG (Koch et al., 2024) to facilitate cross-modal alignment. However, these methods often rely on closed vocabularies and supervised relation annotations. To address these vocabulary constraints, recent approaches have integrated LLMs into scene graph construction. For instance, ConceptGraph (Gu et al., 2024) leverages VLMs and LLMs to infer open-vocabulary categories and relations, while BBQ (Linok et al., 2025) employs

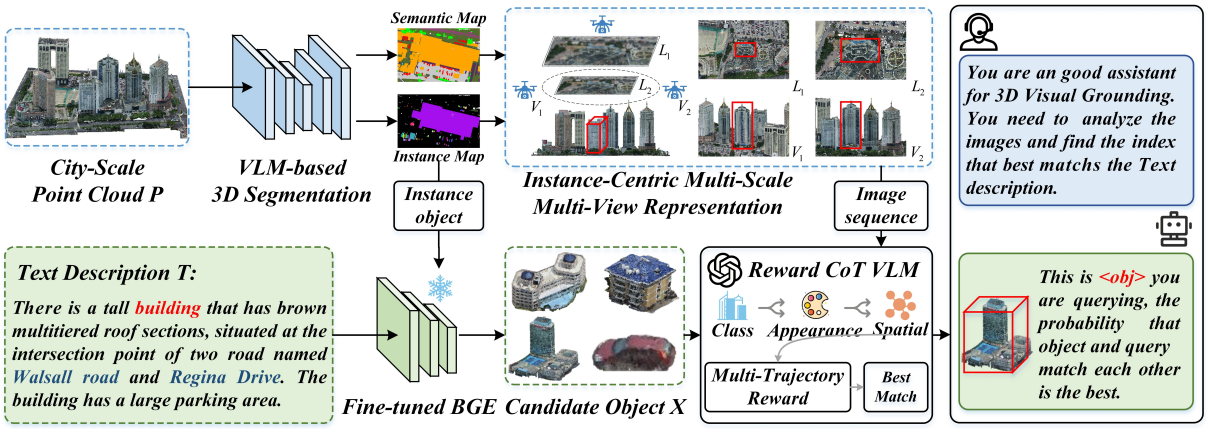

Figure 2: Overview of the proposed CityVG framework. Given a 3D scene point cloud and a textual query, the framework extracts instance-level objects and constructs multi-scale, multi-view representations. Candidate objects are retrieved using a contrastively fine-tuned text encoder and further evaluated by a VLM through reward-based CoT reasoning to localize the target object.

LLMs to generate graph structures for reasoning. However, these methods typically treat LLMs as external generators, constructing graphs without tightly coupling them to the underlying 3D geometric and visual cues.

2.3 Reasoning-Based Inference with Chain-of-Thought

CoT reasoning enhances the inference capabilities of Large Language Models by explicitly generating intermediate reasoning steps (Nguyen et al., 2023; Ho et al., 2023; Hu et al., 2024; Kothapalli et al., 2025). To bolster robustness, prior research has focused on sampling multiple reasoning chains, motivating approaches such as RankCoT (Wu et al., 2025), which identify high-quality chains via ranking mechanisms. Other studies investigate optimizing CoT generation during inference, including DCoT (Puerto et al., 2025), which iteratively refines multiple chains in a single pass, and Soft-CoT (Xu et al., 2025b), which increases efficiency by conducting reasoning in a latent space. (Sun et al., 2025) shows that MLLMs may gradually forget visual evidence over long reasoning chains and proposes Take-along Visual Conditioning to preserve visual information at critical reasoning stages. In parallel, reward modeling (Ryan et al., 2025) has proven effective in guiding inference by aggregating preference signals across candidate outputs, leading to more stable and reliable decision-making. However, existing text-centric methods decouple multi-trajectory exploration from reward modeling, leading to reasoning divergence in complex 3D spaces. We address this by tightly coupling CoT generation with visual-spatial reward aggregation. This transforms CoT into a robust mechanism for spatial disambiguation, dynamically weighing reasoning paths to ensure precise localization.

3 Method

3.1 Overview

Figure 2 illustrates the CityVG framework. Given a city-scale point cloud P and a textual query T, CityVG follows a retrieval-and-reasoning paradigm to localize the target object without manual grounding supervision. The pipeline initiates with VLM-based 3D segmentation designed for open-vocabulary understanding. We synergize a generic 3D instance segmentation model (Mask3D (Schult et al., 2023)) with VLM to perform semantic identification and refinement. This process assigns precise semantic classes to decomposed instances, enabling the system to handle unrestricted vocabulary. Subsequently, for each instance, we construct an instance-centric multi-scale, multi-view representation, generating image sequences that capture diverse perspectives from overhead ($L_{1,2}$) to oblique ($V_{1,2}$) views, as detailed in Appendix A.5. A coarse retrieval stage then prunes the massive search space. Using a Fine-tuned BGE encoder, we align the textual query with these visual representations to filter out irrelevant entities and retrieve a compact set of candidate objects. Finally, a reasoning stage executes fine-grained localization via a Reward CoT VLM. This module processes the candidate image sequences through multi-trajectory Chain-of-Thought inference, aggregating reward

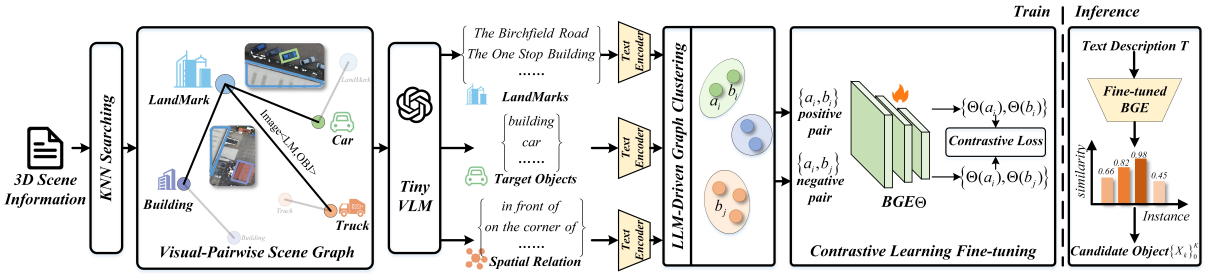

Figure 3: Contrastive fine-tuning strategy of CityVG. During training, instance-level scene graphs are built from object–landmark relationships and converted into relational captions by a VLM. Based on these scene graphs and captions, an LLM-driven clustering strategy constructs positive and negative pairs for contrastive fine-tuning of the BGE encoder. During inference, the textual query is encoded by the fine-tuned BGE to score all instances, producing a set of candidate objects.

signals derived from class, appearance, and spatial consistency to identify the best-matching target robustly.

3.2 Retrieval Stage: Contrastive Fine-Tuning

Visual Pairwise Scene Graph (VPSG). To enable scalable retrieval without relying on predefined labels, we structure the raw point cloud into a VPSG. As shown in Fig. 3, given a set of extracted object instances $\mathcal{O} = \{o_i\}_{i=1}^{N}$, we identify a subset of salient landmarks $\mathcal{L} = \{\ell_k\}_{k=1}^{M}$ ($\mathcal{L} \subseteq \mathcal{O}$) to serve as spatial anchors. We construct a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ comprising two distinct node types: landmark nodes $\mathcal{V}_{\text{land}} = \{\ell_k\}$, which provide stable spatial references, and instance nodes $\mathcal{V}_{\text{inst}} = \{o_i\}$, which represent candidate grounding targets.

Unlike traditional scene graphs (Linok et al., 2025; Gu et al., 2024) that rely on symbolic or purely textual relation predicates, VPSG anchors relational semantics directly in visual observations. Specifically, each node $v_j \in \mathcal{V}$ defines the physical entity through its structural properties: $v_j = (\text{bbox}_j, \text{class}_j)$, where bbox_j denotes the 3D bounding box and class_j is a category label. While nodes provide the structural basis, the semantic richness of the graph stems from visually grounded edges, which link instances to their surrounding landmark context:

$$\mathcal{E} = (o_i \rightarrow \ell_k) \mid \ell_k \in \mathcal{N}(o_i). \tag{1}$$

Each edge is associated with a relation representation $r_{ik} = \Gamma(o_i, \ell_k)$, derived explicitly from paired visual observations of the instance–landmark couple. Specifically, $\Gamma(\cdot)$ takes the image pair of the instance and landmark, along with their relative spatial configuration, and prompts a VLM to generate a free-form textual description of their relationship. By deriving relational semantics from paired visual cues rather than abstract textual priors, VPSG captures nuanced geometric and appearance details, naturally supporting open-vocabulary reasoning and maintaining strict consistency with the underlying 3D geometry.

Notably, landmarks are selected based on semantic distinctiveness and stability. Specifically, we retain instances with explicit and unique semantic identities, such as named buildings, roads, or map-annotated structures, which serve as reliable spatial anchors in the scene graph. These landmarks are not randomly chosen objects, but semantically grounded reference entities that provide stable contextual cues for relational reasoning.

LLM-Driven Graph Clustering (LGC). Building upon the VPSG, we perform LGC to mine reliable pseudo supervision for contrastive fine-tuning (Fig. 4). The procedure initiates with landmark-centered subgraphs, establishing a seed cluster rooted at an anchor instance a_i.

To expand the cluster, we iteratively evaluate candidate instances c_i that share a common landmark context with the anchor. Specifically, we prompt a Large Language Model with the relational descriptions of both the anchor and candidate to compute a semantic consistency score $S(a_i, c_i) \in [0, 10]$ (see Appendix Figure 10). This score serves as a gating metric against a threshold $T_{gate} = 7$: if $S(a_i, c_i) \geq T_{gate}$, the candidate c_i is deemed semantically consistent and merged into the cluster; otherwise, it is excluded. This evaluation repeats until all potential candidates are processed, yielding a collection of semantically coherent clustered scene graphs.

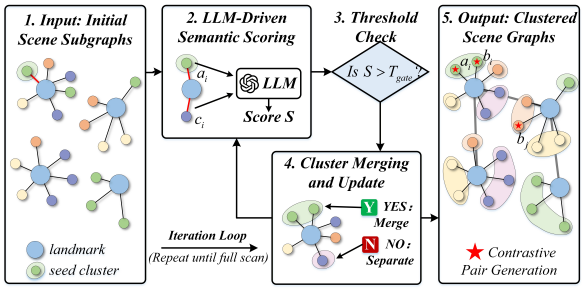

Figure 4: LLM-driven graph clustering for contrastive pair generation. Starting from initial scene subgraphs, an LLM assigns semantic scores to graph components and iteratively merges or separates subgraphs based on a threshold. The resulting clustered scene graphs are used to generate positive and negative pairs for contrastive fine-tuning.

Contrastive Fine-Tuning. The semantically coherent clusters derived from the previous stage serve as the basis for constructing pseudo-supervision. For each anchor instance a_i, we treat co-clustered instances as positive samples and those from disjoint clusters as negatives. Formally, the positive and negative pair sets are defined as:

$$\mathcal{P}^+ = (a_i, b_i), \quad \mathcal{P}^- = (a_i, c_i), \tag{2}$$

where b_i represents an instance within the same cluster as a_i, and c_i denotes an instance from a different cluster.

To adapt the BGE-based text encoder f_θ (BGE-large-en-v1.5 (Xiao et al., 2024)) to the urban domain, we employ a contrastive learning objective driven by composite textual descriptions. For each instance o_i, we synthesize a comprehensive textual representation x_i by integrating its nodal category information (from the VPSG node) with the relational descriptions of its connected landmarks (from the VPSG edges).

Let $\mathbf{t}_a$, $\mathbf{t}_b$, and $\mathbf{t}_c$ denote the encoded embeddings $f_\theta(x)$ for an anchor, a positive, and a negative instance, respectively. We fine-tune f_θ to align the anchor embedding $\mathbf{t}_a$ with its positive counterpart $\mathbf{t}_b$ while distancing it from negatives $\mathbf{t}_c$. The contrastive loss is formulated as:

$$\mathcal{L} = -\frac{1}{|\mathcal{P}^+|} \sum_{(a,b)\in\mathcal{P}^+} \log \frac{\exp(\cos(\mathbf{t}_a, \mathbf{t}_b)/\tau)}{\sum_{(a,c)\in\mathcal{P}} \exp(\cos(\mathbf{t}_a, \mathbf{t}_c)/\tau)}, \tag{3}$$

where $\mathcal{P} = \mathcal{P}^+ \cup \mathcal{P}^-$, $\cos(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter. Optimizing this objective allows the encoder to learn a scene-adaptive embedding space that captures both the object's semantic category and its structural context within the urban layout.

Candidate Retrieval at Inference. During inference, this fine-tuned space facilitates efficient candidate pruning. The input query q is embedded by the adapted encoder as:

$$\mathbf{t}_q = f_\theta(q). \tag{4}$$

Similarly, each object instance o_i in the scene is represented by an embedding $\mathbf{t}_i$, derived from its relational context description. We compute a retrieval score for each instance via cosine similarity:$s_i = \cos(\mathbf{t}_q, \mathbf{t}_i)$. Based on these scores, we select the top-K instances to form a compact candidate set:

$$\mathcal{C} = \{o_i \mid s_i \in \text{Top-}K\}. \tag{5}$$

This reduced set $\mathcal{C}$ is then forwarded to the subsequent reward-based reasoning stage for precise localization, significantly reducing the computational burden compared to exhaustive search.

3.3 Reasoning Stage: Multi-Trajectory Reward-Based CoT

Given the compact candidate set $\mathcal{C}$ retrieved in the previous stage, the reasoning module aims to precisely localize the target object under complex appearance and spatial constraints. Rather than relying on a single reasoning path, CityVG adopts a multi-trajectory reward-based CoT paradigm to improve robustness and reduce ambiguity. The details are summarized in Algorithm 1 and Appendix Fig. 11.

Structured Visual Input. For each candidate $c_i \in \mathcal{C}$, the input to the VLM is a fixed, structured image sequence rather than a single image. Specifically, c_i is represented by a vertically concatenated visual stack consisting of a raw instance-centric image, a landmark context image, and a set of multi-scale and multi-view images $\{l_1, l_2, v_1, v_2\}$ in which the target instance is highlighted with a red bounding box. This image sequence provides consistent visual evidence for all reasoning trajectories associated with c_i.

Parallel Multi-Trajectory Reasoning Hypotheses. Conditioned on the query text q and the structured image sequence c_i, the VLM produces M independent reasoning hypotheses as $\{\pi_i^{(m)}\}_{m=1}^{M}$, each corresponding to a distinct CoT interpretation of the query with respect to candidate c_i. All

Algorithm 1: Multi-Trajectory Reward-Based CoT Reasoning

Input: Query text q; Candidate set $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$;
Number of CoT trajectories M; Weights $\lambda_1, \lambda_2, \lambda_3$.
Output: Index of the grounded target object i^*.
Initialization: $R \leftarrow [\,]$;
(1) Parallel multi-trajectory CoT generation
for *each candidate* $c_i \in \mathcal{C}$ **do**
  $R_i \leftarrow [\,]$;
  in parallel for $m = 1$ ***to*** M **do**
    $\mathrm{CoT}^{(m)} \leftarrow \mathrm{VLMGenerate}(q,\ c_i)$;
    $S_1 \leftarrow \mathrm{EvalCategory}(\mathrm{CoT}^{(m)}_{\mathrm{stage1}})$;
    $S_2 \leftarrow \mathrm{EvalAppearance}(\mathrm{CoT}^{(m)}_{\mathrm{stage2}})$;
    $S_3 \leftarrow \mathrm{EvalSpatial}(\mathrm{CoT}^{(m)}_{\mathrm{stage3}})$;
    $R_i^{(m)} \leftarrow \lambda_1 S_1 + \lambda_2 S_2 + \lambda_3 S_3$;
    Append $R_i^{(m)}$ to R_i;
  end
end
(2) Coupled reward aggregation (Best-of-M)
for $i = 1$ ***to*** K **do**
  $R_i^{\mathrm{final}} \leftarrow \max_m R_i^{(m)}$;
end
(3) Final target selection
$i^* \leftarrow \arg\max_i R_i^{\mathrm{final}}$;
return i^*.

hypotheses are sampled in parallel using a fixed prompt template, ensuring that multi-trajectory reasoning improves robustness without introducing additional inference latency.

Reward-Based Trajectory Evaluation. Each reasoning hypothesis is assessed through a structured reward signal that decomposes inference quality into three interpretable components, yielding category, appearance, and spatial consistency scores $\{S_1, S_2, S_3\}$. These components are combined into a trajectory-level reward $R_i^{(m)}$, which measures how well the hypothesis satisfies the semantic and spatial constraints imposed by the query for candidate c_i.

Coupled Multi-Trajectory Aggregation for Inference. Rather than treating multiple reasoning hypotheses as independent outputs, CityVG explicitly couples hypothesis exploration with reward-based aggregation. For each candidate, the final inference score is obtained via a Best-of-M strategy as:

$$R_i^{\mathrm{final}} = \max_m R_i^{(m)}, \tag{6}$$

which selects the most consistent reasoning hypothesis while suppressing noisy or contradictory ones. The grounded target is then determined by

$$i^* = \arg\max_i R_i^{\mathrm{final}}. \tag{7}$$

By treating CoT trajectories as evaluable reasoning hypotheses and coupling them with reward-based aggregation, the proposed reasoning stage performs robust inference under complex visual and spatial constraints, rather than relying on a single generated explanation.

4 Experiments

4.1 Implementation Details

All experiments are implemented with PyTorch on a cluster of four NVIDIA 5090 GPUs (32 GB), strictly adhering to the official evaluation protocols of CityRefer and CityAnchor. Regarding model instantiation, we utilize the offline Qwen3-VL-8B-Instruct (Bai et al., 2023) for VLM-based 3D segmentation and the lightweight Qwen3-VL-2B-Instruct to construct the VPSG for contrastive fine-tuning; for the reasoning stage, we leverage the Doubao-Seed-1.6 model via its online API. The text encoder BGE (Xiao et al., 2024) is fine-tuned using the SentenceTransformers framework, optimized via a Triplet Loss with cosine distance. We train for 50 epochs with a batch size of 64 and a learning rate of 2×10^{-5} using the AdamW optimizer with a 5% linear warmup. All embeddings are ℓ_2-normalized during both training and inference.

4.2 Qualitative Comparison

Table 1 presents the quantitative results on the CityRefer (Miyanishi et al., 2023) and CityAnchor (Li et al., 2025a) benchmarks. Early indoor-centric methods (e.g., InstanceRefer (Yuan et al., 2021), 3DVG-T (Zhao et al., 2021), EDA (Wu et al., 2023)) struggle to generalize to city-scale environments, with Acc@0.50 consistently falling below 10%. While city-scale baselines like CityRefer and CityAnchor achieve significantly better performance (e.g., CityAnchor reaches 46.86% on CityRefer-NO), they rely heavily on supervised training. In contrast, CityVG achieves competitive performance without any grounding supervision, reaching 40.28% Acc@0.50 on CityRefer-NO and 38.58% Acc@0.50 on CityAnchor-ND, substantially outperforming most supervised methods. Furthermore, when supervised signals are incorporated (CityVG†), our method improves to 50.92% Acc@0.50 on CityRefer-NO, surpassing the state-of-the-art CityAnchor. These results demonstrate that CityVG effectively balances scalability and accuracy, establishing a strong baseline for both

Table 1: Quantitative results on the CityRefer (Miyanishi et al., 2023) and CityAnchor datasets (Li et al., 2025a). "NO" and "ND" denote "Novel Objects" and "Novel Descriptions", respectively. "†" indicates that supervised signals are used during BGE fine-tuning.

Method	Venue	Supervision	CityRefer-NO		CityRefer-ND		CityAnchor-NO		CityAnchor-ND	
			Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
InstanceRefer	ICCV'21	Fully	4.09	3.64	1.93	1.76	1.58	1.35	3.04	2.31
3DVG-T	ICCV'21	Fully	7.73	5.69	9.64	8.12	4.16	2.38	6.25	4.17
EDA	CVPR'23	Fully	6.96	5.53	8.39	5.84	5.15	3.09	7.14	4.29
CityRefer	NeurIPS'23	Fully	8.34	7.47	5.07	3.49	5.73	4.16	6.07	3.95
CityAnchor	ICLR'25	Fully	50.69	46.86	53.17	50.37	41.23	35.11	47.81	43.40
CityVG†	**Ours**	**Fully**	**55.38**	**50.92**	**57.68**	**53.52**	**44.29**	**38.65**	**50.67**	**47.17**
CityVG	**Ours**	**Zero-Shot**	44.13	40.28	44.89	42.86	40.53	35.65	42.64	38.58

Table 2: Ablation study on different component configurations across CFT (LGC, CL) and MRCoT (MT, CoT). CR@K (Candidate Recall@K) measures whether the target is included in the retrieved candidate set at K = 1 and K = 5.

Id	CFT		MRCoT		CR@K		Acc@0.50				
	LGC	CL	MT	CoT	K=1	K=5	Building	Car	Ground	Parking	Overall
(a)	✗	✓	✗	✗	31.58	55.79	43.80	36.53	37.20	31.74	39.31
(b)	✗	✓	✓	✗	31.58	55.79	45.61	38.52	39.15	33.80	41.25
(c)	✗	✓	✗	✓	31.58	55.79	45.09	39.84	38.43	33.16	41.09
(d)	✗	✓	✓	✓	31.58	55.79	48.72	40.84	40.91	34.32	42.83
(e)	✗	✗	✓	✓	–	–	43.05	33.27	34.47	29.66	37.09
(f)	✓	✓	✓	✓	**38.95**	**66.05**	**53.56**	**42.33**	**45.64**	**38.62**	**44.13**

zero-shot and supervised city-scale grounding.

4.3 Qualitative Results

Figure 5 visualizes representative qualitative results on the CityRefer benchmark. As shown, the contrastive fine-tuning stage effectively retrieves a small set of semantically relevant candidates, even when queries involve complex landmarks or subtle appearance attributes. Subsequently, the reasoning stage successfully disambiguates among visually similar instances by jointly evaluating appearance and spatial consistency. These examples validate CityVG's ability to accurately localize targets in crowded urban scenes where traditional single-stage matching often fails.

4.4 Model Analysis

Ablation Studies of Key Modules. Table 2 dissects the contribution of each component within our framework. Replacing LLM-driven graph clustering (LGC) with standard K-Means clustering or removing contrastive learning (CL) leads to a significant drop in candidate recall, confirming the critical role of graph-aware pseudo-supervision in establishing a robust search space. Comparing rows (a)–(d) reveals that enabling either multi-trajectory sampling (MT) or Chain-of-Thought reasoning (CoT) improves overall accuracy, while combining both yields the most substantial gain. The full model achieves 44.13% Acc@0.50, outperforming all ablated variants. This demonstrates that multi-trajectory exploration and reward-based aggregation are complementary mechanisms that jointly enhance robust localization.

Analysis on Key Parameters. Figure 6 analyzes the influence of the candidate size k and the number of reasoning trajectories s. The overall accuracy increases as k grows from 1 to 8 and reaches its maximum at $k = 8$, after which it slightly decreases. Accordingly, we set $k = 8$ as the default candidate size. Similarly, increasing s improves performance, but the gain becomes marginal beyond $s = 3$. To balance robustness and efficiency, we adopt $s = 3$ in our experiments.

Analysis on Model Efficiency. Table 3 compares inference latency across different methods. Despite incorporating multi-trajectory reasoning, CityVG achieves lower inference times than CityAnchor on both benchmarks. This efficiency gain stems from our design choice to restrict heavy reasoning to a

Figure 5: Qualitative results on the CityRefer benchmark. The visualization displays 2D maps derived from city-scale point clouds via top-view and oblique-view projections. Candidate objects retrieved by the Contrastive Fine-Tuning (CFT) stage are highlighted with distinct colored masks. In the textual query, key semantic components are color-coded: the target object is marked in red, appearance attributes in orange, landmarks in blue, and spatial relationships in green. The final grounded object is indicated by a red bounding box.

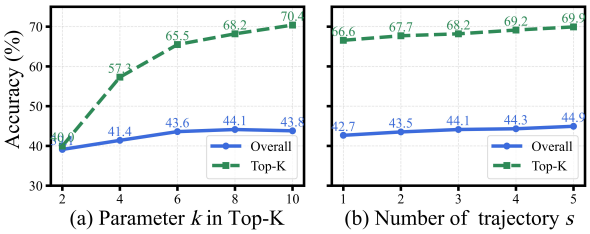

Figure 6: Ablation studies on parameter k and s.

Table 3: Comparison of inference time on CityRefer and CityAnchor datasets. "Time" means the time used in one visual grounding inference.

Method	CityRefer Time (s)	CityAnchor Time (s)
CityRefer	42.27	98.91
CityAnchor	32.45	51.72
CityVG	**29.47**	**49.65**

compact, high-quality candidate set and to sample trajectories in parallel. These results confirm that CityVG improves robustness without incurring significant computational overhead.

5 Conclusion

In this work, we present CityVG, a retrieval-and-reasoning framework for city-scale zero-shot 3D visual grounding. By decoupling candidate pruning from fine-grained inference, it achieves scalable, annotation-free object localization in massive urban point clouds. To support this, we introduce a contrastive fine-tuning strategy anchored by a Visual Pairwise Scene Graph and an LLM-driven clustering mechanism, which jointly mine robust pseudo supervision for efficient retrieval. Furthermore, we design a multi-trajectory reward-based CoT mechanism that tightly couples parallel reasoning with structured reward aggregation for precise localization. Extensive experiments on the CityRefer and CityAnchor benchmarks demonstrate that CityVG achieves strong zero-shot performance and competitive efficiency.

Limitations

Despite its effectiveness, CityVG has limitations. First, while the retrieval stage prunes the search space, the reasoning stage remains dependent on computationally intensive Vision-Language Models, which may scale poorly with a large number of candidates or trajectories. Second, our framework currently focuses on static object-level grounding, overlooking temporal dynamics and multi-object interactions essential for tasks like navigation. Future work will focus on optimizing inference efficiency and extending the framework to model dynamic, long-horizon scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62371343.

References

Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 2019. 3D Scene Graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5664–5673.

Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 2016. 3D semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1534–1543.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. ScanRefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pages 202–221.

Meida Chen, Qingyong Hu, Zifan Yu, Hugues THOMAS, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. 2022. STPLS3D: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. In *The British Machine Vision Conference*.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839.

Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. 2025. Towards Generalizable Vision-Language Robotic Manipulation: A benchmark and llm-guided 3d policy. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 8996–9002.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, and Rama Chellappa. 2024. ConceptGraphs: Open-vocabulary 3d scene graphs for perception and planning. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 5021–5028.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 14852–14882.

Juan Pablo Hourcade, Tristan Braud, Peng Yuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. 2024. All One Needs to Know about Metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *Foundations and Trends in Human–Computer Interaction*, 18(2–3):100–337.

Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. 2022. SensatUrban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, 130(2):316–343.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746.

Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. 2022. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. 2023. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739.

Ue-Hwan Kim, Jin-Man Park, Taek-Jin Song, and Jong-Hwan Kim. 2019. 3-D Scene Graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE Transactions on Cybernetics*, 50(12):4921–4933.

Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. 2024. Open3DSG: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193.

Ping Kong, Ruonan Liu, Zongxia Xie, and Zhibo Pang. 2025. VLN-KHVR: Knowledge-and-history aware visual representation for continuous vision-and-language navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 5236–5243.

Vignesh Kothapalli, Hamed Firooz, and Maziar Sanjabi. 2025. CoT-ICL lab: A synthetic framework for studying chain-of-thought learning from in-context demonstrations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 14620–14642.

Jinpeng Li, Haiping Wang, Yuan Liu, Zhiyang Dou, Yuexin Ma, Sibei Yang, Yuan Li, Wenping Wang, Zhen Dong, and Bisheng Yang. 2025a. CityAnchor: City-scale 3d visual grounding with multi-modality llms. In *International Conference on Learning Representations*.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. 2025b. SeeGround: See and ground for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3707–3717.

Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. 2022. Capturing, reconstructing, and simulating: The urbanscene3d dataset. In *European Conference on Computer Vision*, pages 93–109.

Sergey Linok, Tatiana Zemskova, Svetlana Ladanova, Roman Titkov, Dmitry Yudin, Maxim Monastyrny, and Aleksei Valenkov. 2025. Beyond Bare Queries: Open-vocabulary object grounding with 3d scene graph. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 13582–13589.

Zheng Liu, Jianjun Zhang, Ming Zhang, Runze Ke, Chengcheng Yu, and Ligang Liu. 2026. Unsupervised point cloud reconstruction via recurrent multi-step moving strategy. *IEEE Transactions on Multimedia*, 28:972–984.

Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. 2023. CityRefer: Geography-aware 3d visual grounding dataset on city-scale point cloud data. In *Advances in Neural Information Processing Systems*, volume 36, pages 77758–77770.

Hoang Nguyen, Ye Liu, Chenwei Zhang, Tao Zhang, and Philip Yu. 2023. CoF-CoT: Enhancing large language models with coarse-to-fine chain-of-thought prompting for multi-domain NLU tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 12109–12119.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. 2023. OpenScene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824.

Haritz Puerto, Tilek Chubakov, Xiaodan Zhu, Harish Tayyar Madabushi, and Iryna Gurevych. 2025. Fine-tuning on diverse reasoning chains drives within-inference CoT refinement in LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 3789–3808.

Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. 2024. Multi-branch collaborative learning network for 3d visual grounding. In *European Conference on Computer Vision*, pages 381–398.

Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 8045–8078.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. 2023. Mask3D: Mask transformer for 3d semantic instance segmentation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 8216–8223.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. 2025. Mitigating visual forgetting via take-along visual conditioning for multi-modal long CoT reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 5158–5171.

Ozan Unal, Christos Sakaridis, Suman Saha, and Luc Van Gool. 2024. Four ways to improve verbo-visual fusion for dense 3d visual grounding. In *European Conference on Computer Vision*, pages 196–213.

Yuan Wang, Yali Li, and Shengjin Wang. 2024. G3-LQ: Marrying hyperbolic alignment with explicit semantic-geometric modeling for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13917–13926.

Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. 2023a. Distilling coarse-to-fine semantic matching knowledge for weakly supervised 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2662–2671.

Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, and Lu Sheng. 2023b. VL-SAT: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21560–21569.

Mingyan Wu, Zhenghao Liu, Yukun Yan, Xinze Li, Shi Yu, Zheni Zeng, Yu Gu, and Ge Yu. 2025. RankCoT: Refining knowledge for retrieval-augmented generation through ranking chain-of-thoughts. In *Proceedings of the Annual Meeting of the Association for*

Computational Linguistics, volume 1, pages 12857–12874.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. 2023. EDA: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19231–19242.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2026. Toward Visual Grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2749–2771.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.

Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2025a. VLM-Grounder: A vlm agent for zero-shot 3d visual grounding. In *Proceedings of the Conference on Robot Learning*, volume 270, pages 3961–3985.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025b. SoftCoT: Soft chain-of-thought for efficient reasoning with LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 23336–23351.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. 2024. LLM-Grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 7694–7701.

Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. 2024. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20623–20633.

Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. 2021. InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800.

Jianjun Zhang, Zhipeng Jiang, Qinjun Qiu, and Zheng Liu. 2024. TCFAP-Net: Transformer-based cross-feature fusion and adaptive perception network for large-scale point cloud semantic segmentation. *Pattern Recognition*, 154:110630.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. 2025. MapNav: A novel memory representation via annotated semantic maps for vision-and-language navigation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 13032–13056.

Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 2021. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 2023. 3D-VisTA: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921.

A Appendix

This appendix provides comprehensive supplementary materials to substantiate the findings presented in the main manuscript. We detail the experimental configurations, present extended quantitative and qualitative analyses, and offer in-depth discussions on system design and architectural choices. The appendix is organized as follows:

- Section A.1: Datasets and Evaluation Metrics
- Section A.2: Additional Quantitative Results on Indoor 3DVG
- Section A.3: Qualitative Comparison with Supervised Methods
- Section A.4: Failure Case Analysis
- Section A.5: Candidate Image Sequence Construction
- Section A.6: Systematic Design of Prompt Templates
- Section A.7: Prompt Stage Ablation
- Section A.8: Effect of View Selection
- Section A.9: Effect of Vision-Language Model Choice
- Section A.10: End-to-End 3D Visual Grounding System

A.1 Datasets and Evaluation Metrics

CityRefer Benchmark. CityRefer is a prominent city-scale 3D visual grounding benchmark built upon the large-scale SensatUrban (Hu et al., 2022) point cloud dataset. Covering an urban footprint exceeding 6 km^2, it provides over 35,000 natural language expressions referring to 3D objects, supplemented by more than 5,000 landmark annotations derived from OpenStreetMap to facilitate spatial reasoning. The grounding targets primarily span four semantic categories: *Building*, *Car*, *Ground*, and *Parking*.

CityAnchor Benchmark. CityAnchor is another key benchmark constructed on top of the STPLS3D (Chen et al., 2022) dataset, comprising 25 large-scale outdoor scenes manually annotated with free-form textual descriptions. In total, it contains 1,448 text–object grounding pairs. Compared to CityRefer, CityAnchor presents a more diverse set of object categories—including *Building*, *Vegetation*, *Aircraft*, *Truck*, *Vehicle*, *LightPole*, *Fence*, *StreetSign*, and *Bike*—posing significant challenges for open-vocabulary grounding and category-level generalization.

Evaluation Metrics. We quantify grounding performance using the 3D Intersection over Union (IoU) between the predicted and ground-truth bounding boxes. Formally, given a predicted object $\hat{o}$ and a ground-truth object o, the IoU is defined as:

$$\mathrm{IoU}(\hat{o}, o) = \frac{\mathrm{Vol}(\hat{o} \cap o)}{\mathrm{Vol}(\hat{o} \cup o)}. \tag{8}$$

We report **Acc@0.25** and **Acc@0.50**, which denote the percentage of queries where the IoU exceeds 0.25 and 0.50, respectively.

A.2 Additional Quantitative Results on Indoor 3DVG

Adapting CityVG to Indoor 3DVG. While CityVG is tailored for city-scale environments, its underlying retrieval-and-reasoning paradigm offers versatile adaptability to indoor settings. Given that indoor scenes typically lack dominant landmarks and exhibit lower object density, we recalibrate the Visual Pairwise Scene Graph by pivoting from landmark-centric structures to direct object–object relational graphs based on spatial proximity. Specifically, each object instance connects to a localized set of neighbors, forming pairwise relational contexts that replace the landmark-based hierarchy. In this adapted context, the LLM-driven graph module shifts focus from landmark grouping to pairwise contextual consistency evaluation, assessing whether instances share similar relational patterns to mine pseudo-supervision. Accordingly, we adjust the inference protocol by reducing the candidate pool size to $K = 4$, optimizing for the constrained complexity of indoor layouts. This adaptation enables efficient grounding in indoor scenes by preserving CityVG's structural reasoning capabilities without altering the overall framework.

Quantitative Analysis on ScanRefer. Table 4 presents the zero-shot performance on the ScanRefer validation set, where CityVG is compared against fully supervised methods (Chen et al., 2020; Yuan et al., 2021; Zhao et al., 2021; Jain et al., 2022; Wu et al., 2023; Zhu et al., 2023; Wang et al., 2024; Qian et al., 2024; Unal et al., 2024), a weakly supervised approach (Wang et al.,

Table 4: Comparison of 3DVG performance on the ScanRefer (Chen et al., 2020) validation set. Results are reported for the "Unique" subset (single-target scenes), the "Multiple" subset (scenes with same-class distractors), and the "Overall" split.

Method	Venue	Supervision	Unique		Multiple		Overall	
			Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer	ECCV'20	Fully	67.6	46.2	32.1	21.3	39.0	26.1
InstanceRefer	ICCV'21	Fully	77.5	66.8	31.3	24.8	40.2	32.9
3DVG-T	ICCV'21	Fully	77.2	58.5	38.4	28.7	45.9	34.5
BUTD-DETR	ECCV'22	Fully	84.2	66.3	46.6	35.1	52.2	39.8
EDA	CVPR'23	Fully	85.8	68.6	49.1	37.6	54.6	42.3
3D-VisTA	ICCV'23	Fully	81.6	75.1	43.7	39.1	50.6	45.8
G3-LQ	CVPR'24	Fully	88.6	73.3	50.2	39.7	56.0	44.7
MCLN	ECCV'24	Fully	86.9	72.7	52.0	40.8	57.2	45.7
ConcreteNet	ECCV'24	Fully	86.4	82.1	42.4	38.4	50.6	46.5
WS-3DVG	ICCV'23	Weakly	-	-	-	-	27.4	22.0
LERF	ICCV'23	Zero-Shot	-	-	-	-	4.8	0.9
OpenScene	CVPR'23	Zero-Shot	20.1	13.1	11.1	4.4	13.2	6.5
LLM-G	ICRA'24	Zero-Shot	-	-	-	-	17.1	5.3
ZSVG3D	CVPR'24	Zero-Shot	63.8	58.4	27.7	24.6	36.4	32.7
VLM-Grounder	CoRL'24	Zero-Shot	66.0	29.8	48.3	33.5	51.6	32.8
SeeGround	CVPR'25	Zero-Shot	75.7	68.9	34.0	30.0	44.1	39.4
CityVG	**Ours**	Zero-Shot	**78.4**	**73.9**	**49.6**	**43.2**	**57.8**	**51.3**

2023a), and several zero-shot baselines (Kerr et al., 2023; Peng et al., 2023; Yang et al., 2024; Yuan et al., 2024; Xu et al., 2025a; Li et al., 2025b). Despite being optimized for city-scale scenarios, CityVG demonstrates remarkable cross-domain generalization. In the *Overall* split, it achieves 57.8% Acc@0.25 and 51.3% Acc@0.50, surpassing prior zero-shot baselines like ZSVG3D (36.4% / 32.7%) and VLM-Grounder (51.6% / 32.8%) by a significant margin. Crucially, CityVG exhibits strong robustness on the challenging *Multiple* subset—characterized by same-class distractors—achieving 49.6% Acc@0.25 and 43.2% Acc@0.50. These results confirm that CityVG effectively generalizes beyond its primary urban scope, maintaining strong zero-shot grounding capability in indoor environments.

A.3 Qualitative Comparison with Supervised Methods

Figure 7 compares the supervised baseline CityAnchor with CityVG across representative scenes. As seen in cases (a) and (c), CityAnchor often exhibits spatial drift on large structures despite full supervision, suggesting a failure to capture long-range spatial layouts. In contrast, CityVG produces spatially coherent predictions by leveraging object–landmark relations within the scene graph, accurately localizing targets via their relative positions to landmarks. For visually ambiguous objects in dense environments (cases (b) and (d)), CityAnchor frequently mislocalizes targets due to clutter. CityVG, however, robustly distinguishes targets by employing reward-guided Chain-of-Thought reasoning to enforce appearance and spatial consistency. These results demonstrate that CityVG achieves robust, semantically grounded localization superior to supervised baselines, validating the effectiveness of coupling graph-aware retrieval with multi-path reasoning for zero-shot grounding.

A.4 Failure Case Analysis

Figure 8 illustrates representative failure cases where CityVG generates a *plausible* reasoning trace yet fails to ground the correct instance. Crucially, the reasoning log is generated conditioned on the selected candidate; thus, it may exhibit internal coherence even when the selection is factually incorrect. This phenomenon underscores a critical challenge in prompt-driven reasoning: *explanations can be linguistically consistent without being faithfully grounded in the physical target.*

Failure Mode 1: Challenges in Dense Ordinal Counting. In case (a), the query specifies a gray car positioned "third from the left" within a row of nine vehicles near a landmark. While the model outputs a confident explanation matching the ordinal constraint, the selected candidate is incorrect. Such errors typically arise in highly repetitive environments (e.g., dense parking lots), where identical object appearances and tight spacing make ordinal

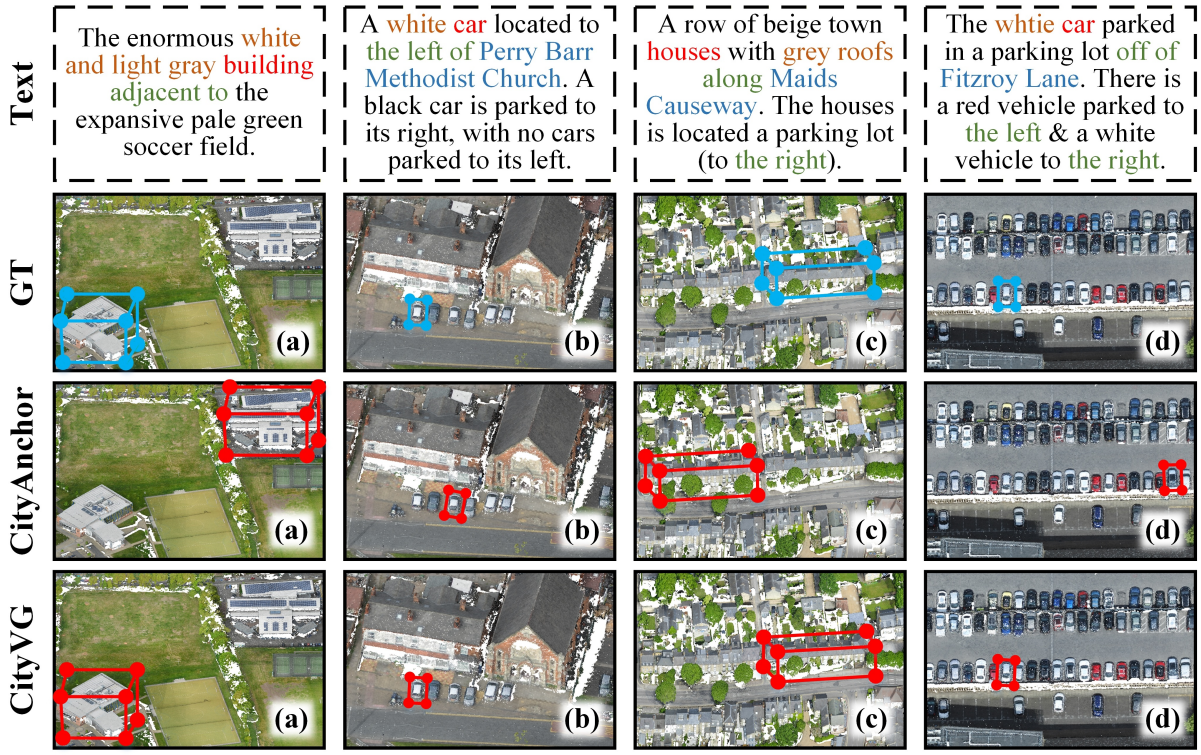

Figure 7: Qualitative comparisons of the supervised method CityAnchor and the proposed framework CityVG. The ground-truth and predicted boxes are displayed in blue and red, respectively.

spatial reasoning (e.g., "third-from-left") highly sensitive to slight viewpoint shifts and ambiguous row definitions (Fig. 8(a)).

Failure Mode 2: Anchor Bias with Partial Attribute Matching. In case (b), the description outlines a building with specific attributes (brown roof, white walls) and context (road adjacency, small lawn, nearby black car). The generated reasoning aligns convincingly with these features, yet the prediction is wrong. This reflects a form of *confirmation bias*: once a candidate aligns with salient primary attributes (e.g., roof color, road adjacency), the model may prematurely commit to it, implicitly treating secondary contextual elements (e.g., lawn, car) as supporting evidence even when they belong to a neighboring structure (Fig. 8(b)).

Failure Mode 3: Ambiguity in Weakly Distinctive Contexts. Case (c) describes an empty rectangular parking lot near Fitzroy Lane, adjacent to trees and a black-roofed building. The model's explanation matches these cues, but the localization is incorrect. This error occurs when multiple regions satisfy the same coarse semantic template (e.g., "empty lot + trees + nearby dark roof"). In such *under-constrained* scenarios, the reasoning remains internally consistent but lacks sufficient discriminative power to distinguish between visually similar regions (Fig. 8(c)).

Failure Mode 4: Multi-Scale Grouping Ambiguity. In case (d), the query targets a red car that is rightmost within a specific lane. The model provides a detailed relational explanation (citing left neighbors and lane indices) but mislocalizes the target. This failure stems from inconsistent definitions of structural groups (e.g., "lanes") across scales. Depending on whether the model groups vehicles by row, structural level, or local cluster, the same spatial arrangement can yield conflicting ordinal indices, allowing the explanation to remain coherent while referencing an incorrect grouping (Fig. 8(d)).

Discussion and Mitigation. These failures suggest that while reward-based CoT enhances interpretability, it does not guarantee factual faithfulness. Future mitigation strategies could include: (1) *Geometric Verification*, introducing explicit consistency checks for critical constraints (e.g., ordinality, adjacency) directly from top-view projections; and (2) *Trajectory Consensus*, penalizing candidates that lack stability across independent reasoning paths to reduce susceptibility to post-hoc rationalization.

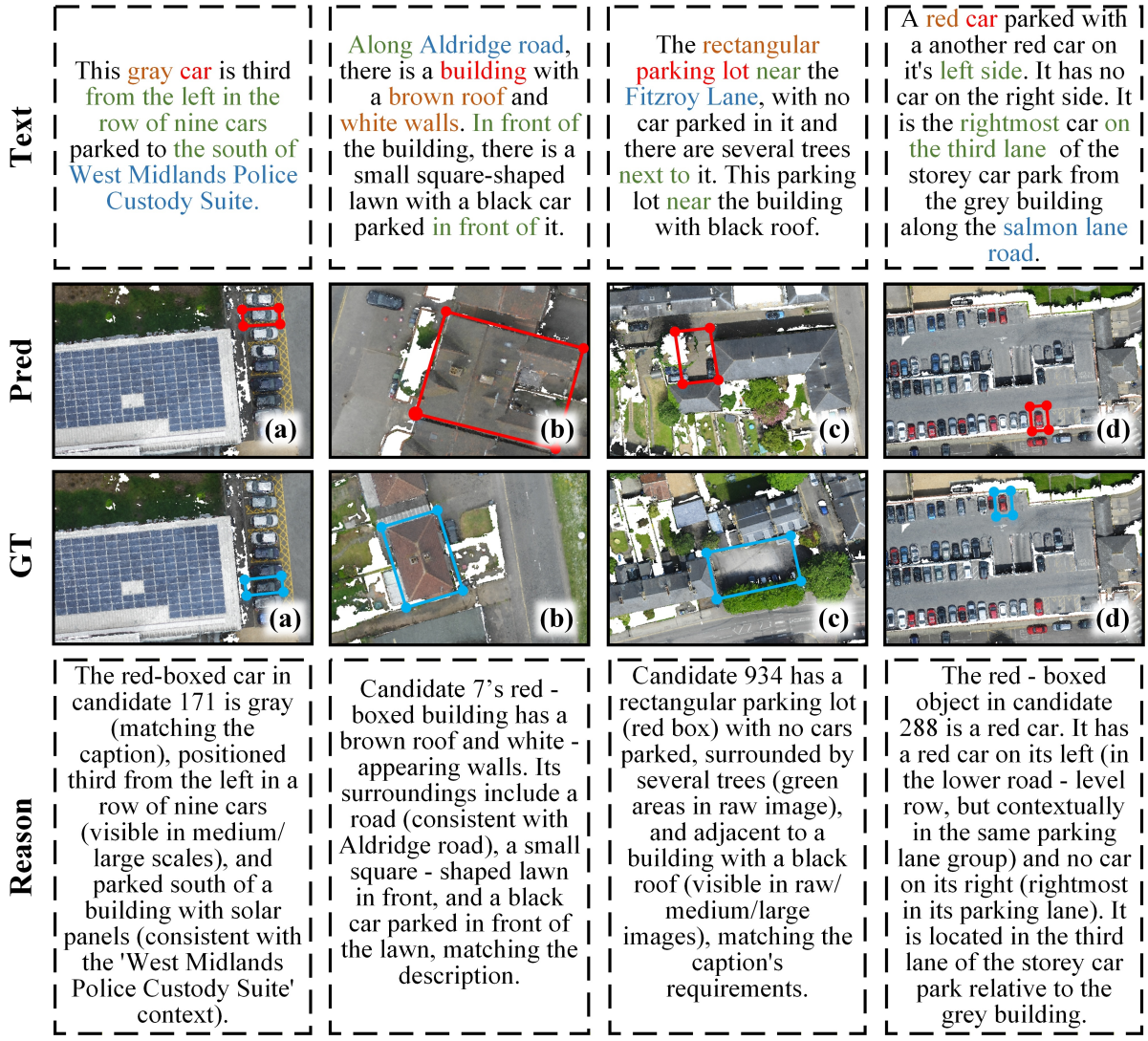

Figure 8: Representative failure cases where CityVG produces plausible reasoning traces but predicts incorrect instances.

A.5 Candidate Image Sequence Construction

To support fine-grained and consistent visual reasoning, we construct an instance-centric image sequence for each candidate object by projecting the city-scale point cloud into multiple 2D views. As illustrated in Fig. 9, we first perform instance segmentation on the input point cloud to obtain object-level instances, from which an instance map is derived.

For each candidate instance, we generate a set of aligned images using instance-centric multi-scale and multi-view projections. Specifically, the image sequence consists of a raw instance-centric top-down image R_1, a landmark context image LM_1, and a set of multi-scale and multi-view images $\{L_1, L_2, V_1, V_2\}$, where the target instance is consistently highlighted with a red bounding box. The landmark image captures the spatial relationship between the target and its surrounding reference structures, while the multi-scale images encode contextual information at different spatial extents. The multi-view images further complement the representation by providing alternative viewing angles to reduce ambiguity caused by occlusion or clutter. All generated images are vertically concatenated in a fixed order to form a unified visual stack. This vertical concatenation enforces a consistent spatial layout across all candidates and serves as shared visual evidence for all reasoning trajectories in the reward-based Chain-of-Thought inference stage. By standardizing the image ordering and layout, the model is encouraged to perform structured and comparable reasoning over appearance, spatial relations, and contextual cues, thereby improving

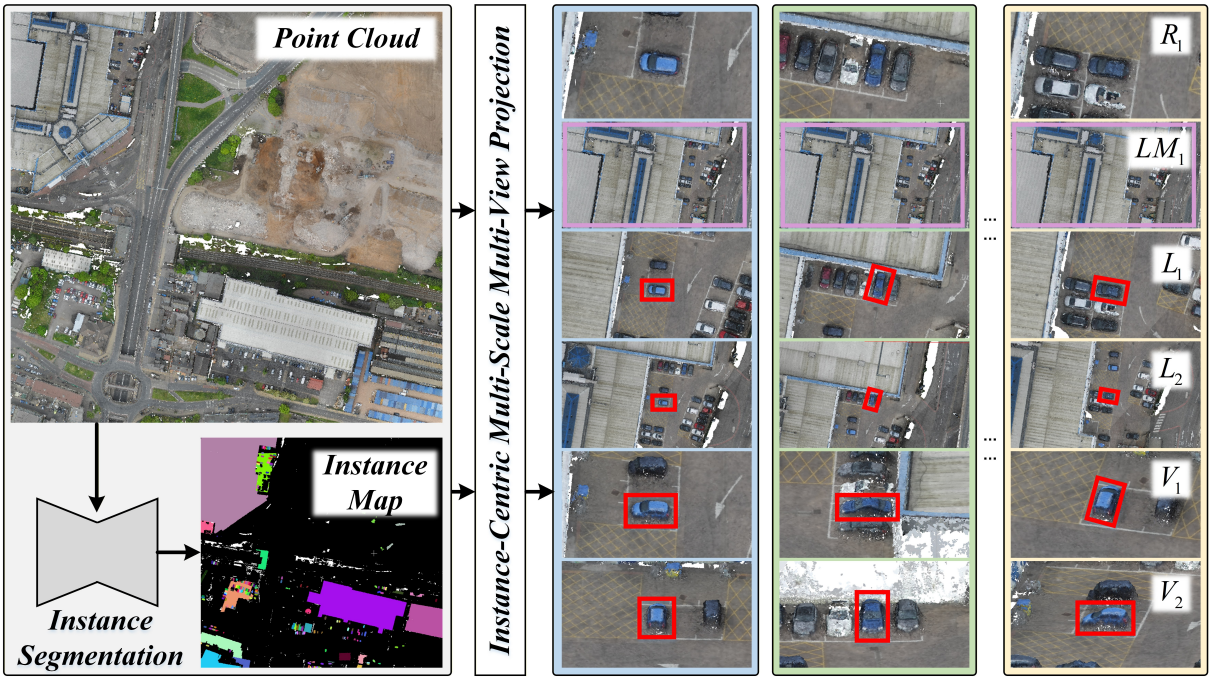

Figure 9: Instance-centric image sequence generation. R_1 denotes the raw instance-centric image, LM_1 represents the landmark context image, and $\{L_1, L_2, V_1, V_2\}$ correspond to multi-scale and multi-view projections. The target instance is highlighted with a red bounding box, and all images are vertically concatenated to form a unified visual stack.

robustness in dense and complex urban scenes.

A.6 Systematic Design of Prompt Templates

Scene Description Generation Prompt. We employ a standardized CityRefer-style prompt to synthesize grounding descriptions for each object–landmark pair (see Fig. 10). The prompt imposes a structured format that explicitly disentangles semantic attributes into three components: object appearance (e.g., color, texture), local positional cues, and landmark-centric spatial relations. Specifically, appearance details are inferred from instance-centric and auxiliary views (R_1, v_1, v_2), while spatial placement is derived from multi-scale top-down and landmark context views (l_1, l_2, LM_1). This view-specific constraint prevents semantic leakage, ensuring that appearance attributes remain distinct from spatial context. Consequently, the generated descriptions function as normalized semantic abstractions of the scene graph. The prompt is applied uniformly across all datasets without manual tuning, facilitating scalable generation and ensuring reproducibility for downstream graph clustering.

Graph Clustering Score Prompt. We utilize a fixed prompt to quantify the semantic compatibility between object descriptions and landmark concepts during graph clustering (see Fig. 10). The prompt instructs the language model to compute a composite similarity score by jointly weighing three factors: category consistency, landmark semantics, and appearance compatibility. This unified scalar signal reflects both intrinsic object attributes and their relational plausibility within the landmark context. Crucially, the prompt enforces a strict JSON output format to ensure deterministic parsing and stable score extraction. This standardized approach allows us to scalably construct high-quality positive and negative clusters for contrastive fine-tuning without relying on explicit supervision.

Multi-Trajectory Reward-Based CoT Prompt. A structured multi-stage prompt guides the fine-grained reasoning process (see Fig. 11). The prompt decomposes grounding into three sequential verification stages: category matching, appearance matching, and spatial relation matching. Each stage imposes strict visual constraints, ensuring that decisions are grounded in relevant evidence. The Vision-Language Model is instructed to output explicit judgments for each stage, which are mapped to binary scores and aggregated into a scalar reward. This reward functions as a self-evaluation signal, allowing the model to rank can-

Prompt for Scene Description Generation

```
You are generating a scene grounding caption for an object named "{object_name}"
and its spatial relation to the landmark "{landmark_name}".

IMAGE INPUTS:
The input consists of a vertically concatenated image sequence with the following views:
1) R1 : raw instance-centric RGB image of the object.
2) LM1 : landmark context image showing the landmark and its surroundings.
3) l1 : medium-scale top-down RGB image
4) l2 : large-scale top-down RGB image
5) v1 : auxiliary close-up view of the object.
6) v2 : auxiliary close-up view of the object.
The target object and the landmark are clearly visible in their corresponding views.
CAPTION STYLE :
Generate a natural descriptive paragraph following this structural pattern as closely as possible:
"The {{color}} {{appearance}}{object_name}{{local_position_phrase}}, {{local_relation_phrase}},
in or near the parking or road area of/behind/next to {landmark_name}{{optional_aerial_phrase}}."
Where:
- {{color}}: the dominant color of the object.
- {{appearance}}: a brief appearance descriptor ,
  inferred jointly from the instance-centric and auxiliary views (R1, v1, and v2).
- {{local_position_phrase}}: describe the local placement of the object within the parking or road layout, inferred from top-down spatial views (l1 and l2),
- {{local_relation_phrase}}: describe relations to nearby visible objects or
  structures using spatial context (l1, l2) and landmark context (LM1),
  e.g. "between a white car and a black car".
- {{optional_aerial_phrase}} (optional): a short aerial reference phrase based on the global layout in l2, such as "near the edge of the map" or
  "with the main road running along the right side".

CONTENT RULES:
1. Sentence 1: describe ONLY the object:
  - object type (following "{object_name}"),
  - main color,
  - appearance attributes (from R1, v1, v2),
  - local position in the parking or road layout,
  - immediately adjacent objects or structures.
2. Sentence 2: describe the relation to "{landmark_name}"
  using a simple spatial phrase derived from LM1 and top-down views,
  such as:
  "in the parking lot of {landmark_name}",
  "behind {landmark_name}",
  "next to {landmark_name}",
  or "near {landmark_name}".
3. Optionally, add a very short phrase about aerial orientation
  using only large-scale context from l2.

STRICT RULES:
- Do NOT mention bounding boxes, colored markers, panels, or image layouts.
- Do NOT refer to the images themselves (e.g., "in the top image").
- Do NOT talk about semantic maps or model views.
- Do NOT hallucinate objects or relations that are not clearly visible.
```

Prompt for Graph Clustering Score

```
"You are a strict semantic similarity evaluator.\n"
    "You must output ONLY a JSON object and NO explanations.\n"
    "The required format is: {\"score\": X} where X is an integer from 0 to 10.\n"
    "\n"
    "Your evaluation must consider THREE factors:\n"
    "1) CATEGORY MATCH: Is the object's category consistent with what usually belongs to the landmark?\n"
    "2) LANDMARK SEMANTICS: Is the object logically part of, associated with, or located at this landmark?\n"
    "3) COLOR / APPEARANCE MATCH: Does the object's color or appearance fit the landmark-related objects?\n"
    "You must integrate all three factors into ONE final similarity score.\n"
    "</system>\n\n"

    "<input>\n"
    f"Landmark Name: {lm}\n"
    f"Object Caption: {cap}\n"
    "</input>\n\n"

    "<task>\n"
    "Evaluate how semantically relevant this Object is to the Landmark.\n"
    "The score must reflect category consistency, landmark relevance, and color/appearance similarity.\n"
    "Return an integer from 0 to 10.\n"
    "STRICTLY output only: {\"score\": X}\n"
    "</task>\n"
)
```

Figure 10: Prompt for Scene Description Generation and Graph-Aware Clustering Scoring.

didates across multiple reasoning trajectories and select the optimal target. The specific reasoning process and result are seen in Fig. 12. Notably, while we employ reward terminology, no parameter updates occur during inference; the reward serves solely as a decision criterion to regularize Chain-of-Thought reasoning and mitigate hallucination. This prompt template is consistently applied across all experiments to ensure fair evaluation.

A.7 Prompt Stage Ablation

Motivation. Our inference protocol utilizes a structured three-stage prompt to sequentially verify *category*, *appearance*, and *spatial relations*, aggregating these judgments into a unified reward. Given the latent reasoning capabilities of modern Vision–Language Models, we investigate the necessity of this explicit decomposition. We perform ablation studies by selectively disabling individual verification stages while holding the candidate set and backbone model constant. Performance is evaluated using Acc@0.50 on the CityRefer benchmark.

Results. Table 5 details the impact of each verification component. Variants (a)–(c) represent the exclusion of category, appearance, and spatial verification, respectively. The removal of any single stage consistently degrades performance relative to the full framework. Notably, omitting spatial reasoning (variant (c)) results in the most significant drop (to 41.30% Acc@0.50), underscoring the critical role of spatial context in urban grounding. The exclusion of appearance (b) or category (a) verification yields moderate yet distinct performance penalties. Crucially, the non-catastrophic nature of these drops confirms the VLM's inherent capacity for implicit reasoning. However, the superior performance of the full three-stage design (44.13% Acc@0.50) demonstrates that explicit, stage-wise verification enforces systematic constraints, ensuring more reliable disambiguation in complex environments populated by visually similar distractors.

A.8 Effect of View Selection

Figure 13 analyzes the complementary roles of top-view and oblique view representations. Top-view projections provide stable global spatial cues, while oblique views capture fine-grained appearance details from ground-level perspectives. Combining these views enables CityVG to jointly reason about spatial layout and visual attributes during inference.

A.9 Effect of Vision–Language Model Choice

Table 6 investigates the impact of the backbone Vision-Language Model (VLM) on reasoning performance. As expected, the models with stronger multi-modal reasoning capabilities yield higher accuracy. Specifically, Doubao-Seed-1.6 achieves

Prompt for Multi-trajectory Reward-based CoT Mechanism

```
You are a 3-Stage CoT Visual Grounding Model.

Each candidate is represented by a vertically concatenated image sequence
in the following fixed order:

1) R1  : raw instance-centric RGB image (no red box)
2) LM1 : landmark context image (purple box)
3) l1  : medium-scale top-down RGB image (red box)
4) l2  : large-scale top-down RGB image (red box)
5) v1  : auxiliary view 1 (red box)
6) v2  : auxiliary view 2 (red box)

The red bounding box always marks the target instance when present.

Your reasoning MUST strictly follow the ordered stages below.

================================================
==
STAGE 1 — CATEGORY MATCHING
================================================
==
- Identify the object category of the target instance.
- You MUST use ONLY the following images:
   • R1  (raw instance-centric image)
   • l1  (medium-scale top-down image)
   • v1  (auxiliary view 1)
- Do NOT use landmark or large-scale context in this stage.
- Do NOT infer category from unrelated background structures.
- Reject candidates whose category is inconsistent with the caption.

================================================
=
STAGE 2 — COLOR & APPEARANCE MATCHING
================================================
=
You must strictly follow the COLOR RULES.

COLOR RULES:
- The object inside the red bounding box is the ONLY target.
- Determine color and appearance using ONLY:
   • R1  (raw instance-centric image)  [PRIMARY]
   • v1, v2 (auxiliary views)          [SECONDARY]
- PRIORITIZE R1 for color estimation.
- ONLY use pixels fully enclosed inside the red bounding box.
- Do NOT use pixels from LM1, l1, or l2 for color judgment.
- Ignore occluded, shadowed, or background pixels.
- Identify dominant color(s), then secondary colors.
- Reject candidates whose color or appearance does NOT match the caption.

================================================
=
STAGE 3 — SPATIAL RELATION MATCHING
================================================
=
After category and appearance are validated, infer spatial relations.

SPATIAL RULES:
- You MUST use ONLY:
   • l1 (medium-scale top-down image)
   • l2 (large-scale top-down image)
- Use these images to reason about:
   • left / right
   • in front of / behind
   • between / next to
   • row or column position (e.g., parking layouts)
   • relative layout of buildings, roads, bridges, or open areas
- Do NOT use instance-centric or auxiliary views in this stage.
- Select the candidate whose spatial configuration best matches the caption.

================================================
======
SELF-REWARD CHECKER
================================================
======
CatScore    = 1 if category matches caption, else 0
ColorScore  = 1 if color/appearance matches caption, else 0
SpatialScore = 1 if spatial relations match caption, else 0

Reward = (CatScore + ColorScore + SpatialScore) / 3

You MUST compute Reward strictly according to the above rules.

================================================
======
OUTPUT FORMAT (STRICT JSON)
================================================
======
Return ONLY one JSON object, no extra text:

{
  "best_id": "<one of [{id_list_str}]>",
  "reason": "<explain reasoning for Stage 1, Stage 2, and Stage 3>",
  "reward": <float between 0 and 1>
}
```

Figure 11: Prompt for Multi-Trajectory Reward-Based Chain-of-Thought Reasoning.

Table 5: Prompt-stage ablation on CityRefer (Acc@0.50). "Category", "Appearance", and "Spatial" denote whether category verification, appearance verification, and spatial relation verification are explicitly enforced in the prompt.

Prompt Variant	Category	Appearance	Spatial	Acc@0.50
(a)	✗	✓	✓	43.05
(b)	✓	✗	✓	42.22
(c)	✓	✓	✗	41.30
CityVG	✓	✓	✓	**44.13**

Table 6: Ablation study of the VLM used for inference.

VLM	Acc@0.50
Qwen2-VL-72B	33.25
GPT-4o	40.34
Doubao-1.5-vision-pro	41.54
Qwen3-VL-Plus	43.07
Doubao-Seed-1.6	**44.13**

the best performance (44.13% Acc@0.50), slightly outperforming Qwen3-VL-Plus (43.07%) and significantly surpassing Qwen2-VL-72B (33.25%). Notably, even general-purpose models like GPT-4o achieve competitive results (40.34%), confirming the robustness of our reward-based Chain-of-Thought mechanism. These results demonstrate that while CityVG benefits from advanced VLMs, its core framework is model-agnostic and consistently effective across diverse architectures.

A.10 End-to-End 3D Visual Grounding System

We present an end-to-end 3D visual grounding system built upon the CityVG framework to demonstrate its practical applicability in real-world scenarios. As illustrated in Fig. 14, the system processes a raw city-scale point cloud and a natural language query through a unified pipeline to localize target objects automatically. Upon selecting a scene in the interactive 3D viewer, users provide a free-form description specifying appearance attributes and spatial relations. The system then triggers the annotation-free CityVG pipeline: first, the retrieval module aligns the query with urban scene graph representations to isolate high-probability candidates; subsequently, the reasoning module executes reward-based multi-path Chain-of-Thought inference to ground the target object. The final

Reasoning Process and Result of Multi-trajectory Reward-based CoT Mechanism


```
{
  "best_id": "128",
  "reason": "Stage1: The object is a car (category matches caption's 'red car'). Stage2: The car's color is red (matches caption). Stage3: It is at the end of the first row against the building, between two darker cars, with a marked area (yellow hatchmarks) in front (spatial relations match).",
  "reward": 1.0,
  "all_trajs": [
   {
    "traj_id": 0,
    "best_id": "128",
    "reward": 1.0,
    "reason": "Stage1: The object is a car (category matches caption's 'red car'). Stage2: The car's color is red (matches caption). Stage3: It is at the end of the first row against the building, between two darker cars, with a marked area (yellow hatchmarks) in front (spatial relations match)."
   },
   {
    "traj_id": 1,
    "best_id": "128",
    "reward": 1.0,
    "reason": "Stage1: Category matches (red car, type car). Stage2: Color matches (red). Stage3: Spatial relations match (end of first row against building, between darker cars, yellow hatchmarks in front)."
   },
   {
    "traj_id": 2,
    "best_id": "128",
    "reward": 1.0,
    "reason": "Stage1: Object category is car (matches caption's 'red car'). Stage2: Color is red (matches caption). Stage3: Spatial - at end of first row against building, between darker cars, with yellow hatchmarked area (no-parking) in front (matches layout)."
   }
  ],
  "R_final": 1.0,
  "low_confidence": false,
  "caption": "At the end of the first row of cars parked against the building from the entry direction of the parking area beside the One Stop, there is a red car. It is between two darker colored cars and there is a marked out area with yellow hatchmarks in front of it.",
  "candidates": ["139","117","128","137", "68","136","129","24","42","130"],
  "GTID": "128"
 },
```

```
{
  "best_id": "65",
  "reason": "Stage 1: The object in the red box is a car (category matches 'blue car' in caption). Stage 2: The car's color is blue (matches raw image pixels). Stage 3: The spatial layout shows a small parking area behind a building, with a black car (dark-colored) adjacent to the blue car and a white van at the end, aligning with the caption's description of the parking area off Aston Lane behind the Jobcentre Plus building. All three scores (Cat, Color, Spatial) are 1, so reward is 1.0.",
  "reward": 1.0,
  "all_trajs": [
   {
    "traj_id": 0,
    "best_id": "65",
    "reward": 1.0,
    "reason": "Stage 1: The object in the red box is a car (category matches 'blue car' in caption). Stage 2: The car's color is blue (matches raw image pixels). Stage 3: The spatial layout shows a small parking area behind a building, with a black car (dark-colored) adjacent to the blue car and a white van at the end, aligning with the caption's description of the parking area off Aston Lane behind the Jobcentre Plus building. All three scores (Cat, Color, Spatial) are 1, so reward is 1.0."
   },
   {
    "traj_id": 1,
    "best_id": "65",
    "reward": 1.0,
    "reason": "Stage1: The object is a car (category matches 'blue car'). Stage2: The car inside the red box is blue (matches 'blue'). Stage3: The blue car is in a small parking area behind a building, with a black car to its left and a white van to its right (matches spatial relations in the caption)."
   },
   {
    "traj_id": 2,
    "best_id": "65",
    "reward": 1.0,
    "reason": "Stage1: Category is a car (sedan), matching the caption's 'blue car'. Stage2: Color is blue, confirmed by raw and red-boxed images. Stage3: Spatial layout matches: small parking area behind a building (consistent with Jobcentre Plus), blue car with black car (left) and white van (right), aligning with 'black car on one side and white van on the other'."
   }
  ],
  "R_final": 1.0,
  "low_confidence": false,
  "caption": "A blue car is parked in a small parking area off Aston Lane behind the Jobcentre Plus building, with a black car on one side and a white van on the other side.",
  "candidates": ["65","67","93","120","69","73","182","75","95","49" ],
  "GTID": "65"
 }
```


Figure 12: Reasoning Process of Multi-trajectory Reward-based CoT Mechanism.

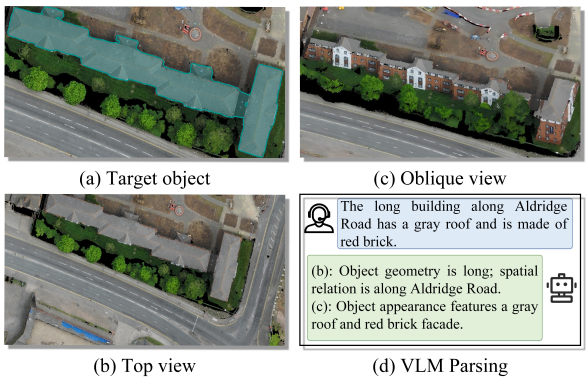

Figure 13: Complementary roles of top-view and oblique-view representations.

output highlights the predicted 3D bounding box within the scene, accompanied by an interpretable reasoning log detailing the selection logic. This transparent design facilitates user inspection while maintaining fully automated inference. Overall, this demonstration validates CityVG's capability to seamlessly integrate scalable retrieval and robust reasoning into a unified, zero-shot system, underscoring its potential for practical city-scale applications.

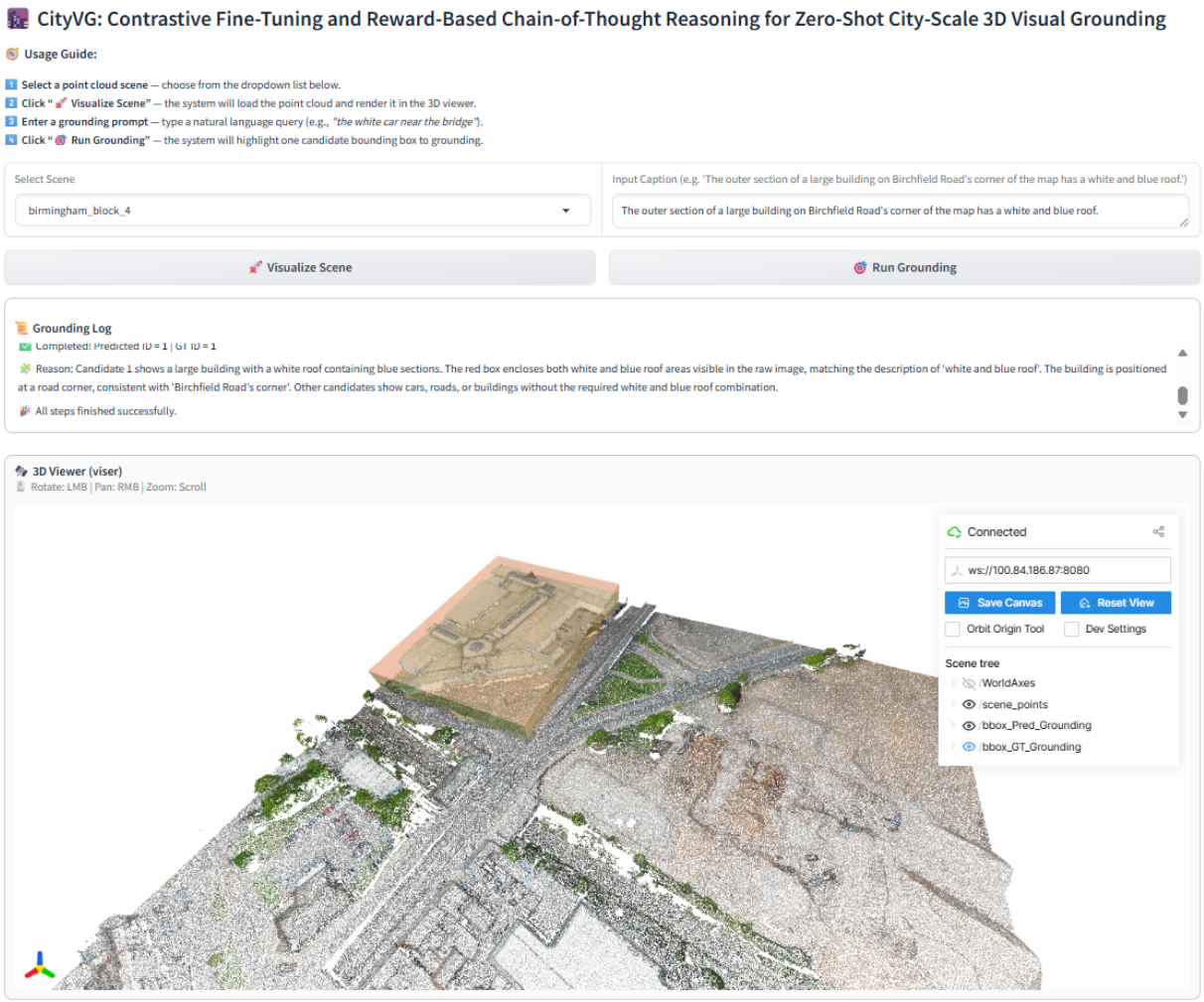

Figure 14: End-to-End demonstration of the CityVG 3D visual grounding system.