Abstract the Layout, Focus the Detail: A Dual-Granularity Representation Framework for Zero-Shot 3D Visual Grounding

Zeyuan Lin[1,2], Hanxuan Li[1,2], Chen He[1,2], Ruiping Wang[1,2(✉)], Zhaoxiang Liu[3,4], and Xilin Chen[1,2]

[1] Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Data Science & Artificial Intelligence Research Institute, China Unicom, Beijing 100033, China
[4] Unicom Data Intelligence, China Unicom, Beijing 100033, China
{zeyuan.lin, hanxuan.li}@vipl.ict.ac.cn, {hechen,wangruiping,xlchen}@ict.ac.cn, liuzx178@chinaunicom.cn

Abstract. Zero-shot 3D Visual Grounding (3DVG) has gained increasing attention for its ability to bypass costly 3D annotations by leveraging pre-trained LLMs and VLMs. However, existing symbol-based and view-based paradigms often fall short in complex spatial reasoning, as they tend to reconstruct global spatial understanding from local evidence, making it difficult to resolve long-range spatial dependencies and mitigate physical occlusions. To address this, we propose a dual-granularity representation framework for zero-shot 3DVG. Our framework abstracts the 3D scene into two representations: a macro-level Semantic Spatial Layout that provides a clear reasoning medium for spatial relationships, and micro-level Visual Detail Patches that focus on the objects' high-fidelity appearance. To make the representation VLM-friendly, we introduce a Query-Guided Scene Filtering module to parse the perception requirements and filter the scene, and we finally conduct Joint Spatial-Visual Reasoning over the representations. Extensive experiments demonstrate that our framework outperforms state-of-the-art zero-shot methods on the ScanRefer and Nr3D benchmarks, showing advantages in resolving long-range and compositional spatial relations. Furthermore, preliminary implementation in real-world scenarios suggests its potential for embodied AI applications.

Keywords: 3D Visual Grounding · Zero-shot Learning · Spatial Reasoning

1 Introduction

3D Visual Grounding (3DVG) aims to establish fine-grained alignment between natural language queries and specific instances within a 3D physical scene, serving as a fundamental ability for embodied AI systems. Over the past few years,

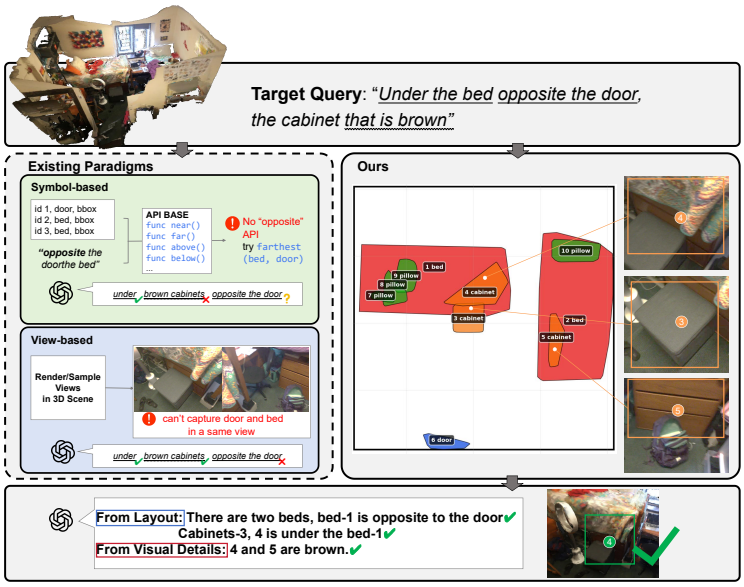

Fig. 1: Comparison of existing and our method for 3D Visual Grounding. Compared to (a) *Symbol-based* methods and (b) *View-based* methods. Our method decouples the scene into a macro-level layout to provide a clear medium for spatial reasoning and micro-level visual details for visual verification, enabling a more robust localization.

fully-supervised paradigms [1, 3, 9, 10, 15, 20, 26, 27, 30] have achieved impressive progress. However, they remain constrained by the prohibitive cost of collecting large-scale 3D-language annotations, which restricts their scalability to open-world environments. To alleviate this dependency on labor-intensive 3D annotations, research has explored zero-shot 3DVG by leveraging pre-trained LLMs and VLMs [6, 12, 14, 16, 23, 24, 28, 29], capitalizing on their extensive open-world knowledge and general reasoning capabilities.

Since native 3D data falls outside the perception scope of LLMs and VLMs, which are primarily optimized for textual and 2D visual inputs, a central challenge lies in reformulating the unstructured 3D space into model-compatible proxies. Existing methods typically follow two paradigms: symbol-based and view-based, as shown in Fig. 1. *Symbol-based methods* [6,28,29] typically abstract 3D scenes into object-level descriptions and coordinates, employing predefined symbolic functions for spatial relation prediction integrated with LLM-based reasoning. A major bottleneck lies in the difficulty of encoding spatial relationships into rigid mathematical rules. For instance, evaluating relations like "opposite" solely through coordinate logic may lose useful visual context and become sensitive to how the relation is formalized. *View-based methods* [12, 14, 23] typically represent the 3D scene through camera views. This effectively exploits the VLM's strong 2D visual reasoning capabilities and demonstrates strong 2D visual understanding. However, local views provide only partial, occluded observations of

the scene. When facing queries with long-range spatial dependencies (e.g., locate the door behind the observer when facing the bed) or complex spatial arrangements, the restricted field-of-view and the inevitable mutual occlusions among 3D objects limit the accuracy of spatial reasoning.

The aforementioned paradigms generally reconstruct global spatial understanding from local evidence, aggregating pairwise symbolic predicates in coordinate space or partial visual observations in camera views. Such bottom-up style reasoning can be brittle when queries require an integrated understanding of room-level layouts, compositional spatial relations, or long-range spatial dependencies. This inspires us to reformulate the scene and build a more robust representation that complements a holistic understanding of the room layout with fine-grained localized views. Such a synergistic design naturally aligns with the classical HCI principle: "*Overview first, zoom and filter, then details-on-demand*" [22].

Based on this observation, we propose a dual-granularity representation framework for zero-shot 3DVG. In this framework, physical instances are explicitly abstracted into two representations: a macro-level **Semantic Spatial Layout** (*Overview first*) that provides a clear reasoning medium for spatial relations without visual distraction, and micro-level **Visual Detail Patches** (*details-on-demand*) that focus exclusively on the unoccluded, high-fidelity appearance of individual objects, liberated from capturing 3D spatial context in 2D view, and provide a verification for Semantic Spatial Layout. To make the representation VLM-friendly, we introduce a **Query-Guided Scene Filtering** module (*zoom and filter*) to explicitly parse the perception requirements and screen out irrelevant 3D background. Finally, the VLM is prompted to perform **Joint Spatial-Visual Reasoning** over the representations. By abstracting the layout and focusing the details, our framework effectively mitigates spatial hallucinations and minimizes visual noise.

Extensive experiments on ScanRefer and Nr3D show that our framework achieves strong zero-shot performance, outperforming existing zero-shot methods on both benchmarks. Compared with agent-based methods, our framework achieves superior results with remarkable efficiency. Furthermore, preliminary deployments in real-world scenarios suggest its potential for embodied AI applications such as object-driven navigation.

2 Related Work

2.1 Supervised 3D Visual Grounding

3D visual grounding (3DVG) is a fundamental 3D vision-language task aiming to accurately localize target objects in 3D scenes based on natural language queries [1, 3]. Traditional methods predominantly rely on fully-supervised paradigms [1,3,9,10,15,20,26,27,30]. These approaches are generally categorized into two-stage and one-stage frameworks. Two-stage methods first extract object proposals using 3D detectors and then perform semantic matching. To handle

complex indoor scenarios, researchers have explored various inter-object relationship modeling strategies. These range from non-relational matching [3] to explicit relational modeling utilizing Graph Neural Networks [1,9]. Furthermore, implicit contextual modeling via Transformer attention mechanisms [20, 26, 30] has also been widely adopted to fuse multi-level contextual features. Conversely, one-stage methods [10,15,27] bypass proposal generation and directly fuse modalities to regress target bounding boxes, improving computational efficiency. Recently, to overcome the limitations of domain-specific architectures, methods based on large-scale 3D vision-language pre-training [25,33] have been proposed to establish universal 3D-text representations through massive self-supervised data. Despite their impressive benchmark performance, these supervised and pre-trained models are still dependent on expensive, fine-grained 3D-text annotations, which limits the scalability.

2.2 Zero-shot 3D Visual Grounding

Recent research explores zero-shot 3DVG as a promising paradigm. Some methods primarily utilize Large Language Models (LLMs) to build agent workflows, decomposing the localization task into manageable sub-steps. For instance, LLM-Grounder [24] utilizes open-vocabulary perception tools [18] to extract scene objects and employs LLMs for text-based relational reasoning. ZSVG3D [29] and Transcrib3D [6] guide LLMs to generate visual programs or symbolic representations to facilitate spatial reasoning. Similarly, CSVG [28] formulates the zero-shot grounding task as a Constraint Satisfaction Problem (CSP) to enable global symbolic reasoning, while LaSP [16] introduces an automated pipeline to evaluate and optimize LLM-generated spatial reasoning codes. Although these methods excel at modeling logical and spatial relations, their reliance on coordinate-abstracted scene representations constrains their ability to handle queries that demand fine-grained visual attributes.

Under another paradigm, some studies integrate Vision-Language Models (VLMs) capable of directly processing 2D observational inputs. SeeGround [14] incorporates local rendered images to provide additional visual contexts for VLMs; VLM-Grounder [23] directly utilizes 2D video sequences by stitching multiple frames to bypass complex 3D representations. SPAZER [12] further designs a spatial-semantic progressive reasoning pipeline, which attempts to introduce the global scene context through a holistic rendered view. While this global rendering indeed provides a broader perspective and achieves significant improvement, it remains to project the complex 3D environment onto specific viewpoints, leaving the potential for a unified representation that transcends individual observational views underexplored.

2.3 Vision-Language Models for 3D Scene Understanding

There is a growing trend of leveraging Vision-Language Models to facilitate broader 3D scene understanding. Early explorations primarily focused on lifting robust 2D features from VLMs into 3D spaces via knowledge distillation

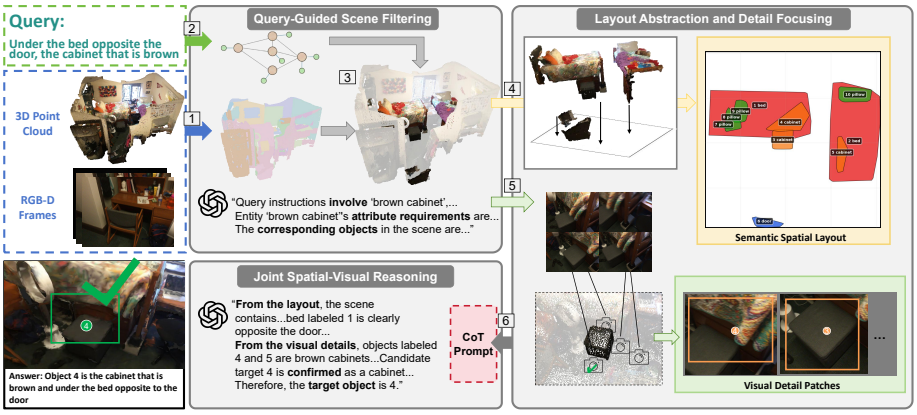

Fig. 2: Overview of our dual-granularity representation framework. The framework consists of **Query-Guided Scene Filtering** (Sec. 3.2), **Layout Abstraction and Detail Focusing** (Sec. 3.3), and **Joint Spatial-Visual Reasoning** (Sec. 3.4)

or multi-view projection, as demonstrated by OpenScene [18] and LERF [13], enabling open-vocabulary 3D segmentation and retrieval. Simultaneously, some research [7] employs VLMs to construct structured cognitive representations or perform direct scene reasoning. To further bridge the modal gap, Chat-Scene [8] introduces object identifiers to translate 3D scene structures into sequential text tokens, enabling LLMs to process 3D environments through a unified question-answering format. Moreover, recent approaches like Video-3D LLMs and GPT4Scene [19] explore understanding 3D scenes directly from 2D video sequences. These diverse explorations validate the profound adaptability of VLMs in 3D scene understanding.

3 Method

3.1 Overview

3D Visual Grounding (3DVG) aims to locate a target object o^* within a 3D scene $\mathcal{S}$, given a natural language query $\mathcal{Q}$. Formally, the objective is to learn a mapping $f : (\mathcal{S}, \mathcal{Q}) \rightarrow \mathcal{B}^*$, where $\mathcal{B}^*$ is the 3D bounding box of o^*.

As illustrated in Fig. 2, we propose a dual-granularity representation framework for zero-shot 3D Visual Grounding. To bridge the modality gap without inducing information overload, our framework formulates the grounding task through three cohesive modules:

1) **Query-Guided Scene Filtering** (Sec. 3.2), which acts as a cognitive sieve to extract query-relevant entities, proactively filtering the cluttered 3D environment into a compact, high-signal sub-scene;

2) **Layout Abstraction and Detail Focusing** (Sec. 3.3), which translates this filtered sub-scene into a VLM-friendly, dual-granularity representation—a

macro-level Semantic Spatial Layout (SSL) for topological memory and micro-level Visual Detail Patches (VDP) for appearance verification;

3) **Joint Spatial-Visual Reasoning** (Sec. 3.4), which integrates these decoupled representations via a natural "Spatial-First, Visual-Second" workflow, guiding the VLM to robustly deduce the final target without relying on complex reasoning rules.

3.2 Query-Guided Scene Filtering

Natural language queries usually implicitly define a specific semantic context, dictating which objects and attributes are essential for localization. As directly exposing the entire unconstrained 3D scene to downstream Vision-Language Models (VLMs) inevitably overwhelms them with redundant visual and geometric noise, we utilize the query information to proactively filter the scene.

We design the Query-Guided Scene Filtering module to act as a crucial cognitive sieve. Its primary goal is to explicitly extract the query-relevant semantic context and filter out distracting background instances before any spatial layout construction or visual verification occurs. By doing so, we construct a compact, high-signal sub-scene. This proactive filtering significantly reduces the cognitive load for subsequent reasoning, ensuring that the abstracted representations purely reflect the essential structure demanded by the query. Conceptually, this process is decoupled into three sequential steps: scene initialization, query context analysis, and robust semantic alignment.

Scene Initialization Given a continuous 3D scene $\mathcal{S}$, we first utilize an off-the-shelf 3D instance segmentation network to extract a set of object proposals $\mathcal{P} = \{p_1, p_2, \ldots, p_K\}$, following the previous setting [14, 29]. However, relying solely on closed-set pre-trained categories is inadequate for the open-vocabulary nature of 3DVG.

To bridge this gap, we employ a Vision-Language Model (VLM) to assign semantic labels to these proposals as [32]. Nevertheless, directly prompting a VLM with 2D projections often leads to spatial ambiguities (e.g., misclassifying books on a shelf as the whole bookshelf due to local perspectives). To alleviate this, we inject the intrinsic 3D proposal categories as a *soft prior*.

Specifically, for each proposal p_k, we heuristically select its optimal 2D frame v_k^* (the same process as detailed in Sec. 3.3), then project the 3D mask of p_k onto v_k^* to generate a visually prompted image (e.g., highlighting the target region in green). Both the original RGB frame and the mask-annotated frame are fed into the VLM, accompanied by a text prompt that explicitly states the prior 3D category of p_k. Through a localized Chain-of-Thought, the VLM performs prior-guided open-vocabulary perception and assigns a robust text label l_k to each p_k. The final perceived 3D instances are denoted as $\mathcal{I} = \{o_1, o_2, \ldots, o_K\}$ with their corresponding semantic labels $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$.

Query Context Analysis Once the physical scene is semantically initialized, we utilize a Large Language Model (LLM) to bridge the linguistic query $\mathcal{Q}$ with the scene instances $\mathcal{I}$. The first conceptual step is to dissect the semantic context of $\mathcal{Q}$, extracting the hypothetical entities required for localization and explicitly decoupling the perception granularity demanded by each entity.

Specifically, we utilize the LLM to parse $\mathcal{Q}$ into a set of distinct entities $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$. Each entity e_i is parameterized as $(d_i, \mathbf{a}_i)$, where d_i is the textual description and $\mathbf{a}_i = \{\tau_{tgt}, \tau_{vis}, \tau_{hgt}, \tau_{cam}, \tau_{pos}\}$ are boolean state flags indicating whether e_i is the target candidate, requires fine-grained visual verification, demands vertical height context, or involves specific viewpoint conditions, respectively.

By extracting these condition flags, we ensure that computationally expensive visual verifications are triggered strictly on-demand in subsequent stages, significantly reducing token consumption and minimizing visual noise.

Robust Semantic Alignment Following the extraction of linguistic entities $\mathcal{E}$, the subsequent conceptual step is to establish their physical correspondence with the initialized scene instances $\mathcal{I}$. To maximize inference efficiency, this alignment is practically executed concurrently with the context analysis within a single LLM reasoning pass.

Due to the inherent flexibility of natural language and the significant intra-class variations of objects in real-world 3D environments, a rigid string-matching algorithm is insufficient. Instead, we perform a robust soft mapping strategy. Each parsed entity e_i is evaluated and mapped to a semantic subset $L_i \subseteq \mathcal{C}_{scene}$, where $\mathcal{C}_{scene}$ denotes the set of unique labels in the scene $\mathcal{L}$. By leveraging synonyms, hypernym/hyponym reasoning, and functional equivalence, this generalized mapping successfully bridges the vocabulary gap. Ultimately, all instances in $\mathcal{S}$ whose assigned labels fall into L_i are retrieved to form the candidate set for e_i: $\mathcal{I}_i = \{o_k \in \mathcal{I} \mid l_k \in L_i\}$.

Through these parsing and alignment steps, the vast and cluttered 3D scene is effectively filtered into a task-aware, high-signal context. This targeted filtering lays a mathematically and semantically clean structural foundation for the downstream scene abstraction.

3.3 Layout Abstraction and Detail Focusing

3D scenes inherently possess a hierarchical information structure, encompassing both macro-level spatial relations and micro-level visual attributes. To equip the VLM with a comprehensive 3D cognition without inducing information overload, we abstract the filtered sub-scene into a dual-granularity representation.

We construct an explicit **Semantic Spatial Layout (SSL)** to provide a global spatial memory and serve as a reasoning medium for global spatial relationships, while simultaneously extracting **Visual Detail Patches (VDP)** to exclusively capture the unoccluded appearance of individual objects. This

parallel abstraction naturally aligns with the hierarchical nature of 3D environments, enabling the VLM to process spatial and visual cues in their suitable representation forms.

Macro-level Semantic Spatial Layout (SSL) To endow the VLM with an intuitive global spatial memory, we abstract the filtered candidate instances $\hat{\mathcal{I}} = \bigcup_{i=1}^{N} \mathcal{I}_i$ into a highly condensed layout representation.

Specifically, we project the 3D geometries of all instances in $\hat{\mathcal{I}}$ onto the ground plane (XY-plane). Let each instance $o_k \in \hat{\mathcal{I}}$ be represented by its raw 3D point cloud $P_k \in \mathbb{R}^{M_k \times 3}$. To abstract the macro-level layout and reduce geometric noise, we perform an orthogonal projection onto the XY-plane and compute the 2D convex hull $\mathcal{H}_k$: $\mathcal{H}_k = \text{Convex}(P_k^{xy})$. These geometric abstractions $\mathcal{H}_k$ are then rendered onto a global canvas I_{SSL}. Each instance is assigned a unique categorical color and a clear numeric ID overlay. To provide absolute spatial references and alleviate language model hallucinations on distance, a coordinate grid and peripheral axes are augmented to the layout. Crucially, since VLMs often struggle with zero-shot mental rotation, we introduce a *Query-Adaptive Viewpoint Alignment* mechanism. Leveraging the parsed viewpoint conditions (τ_{cam} and τ_{pos}), we identify the reference anchor and target direction implied by the query. We then rotate the layout canvas so that the implied observer's perspective (e.g., "facing the door") is aligned with the canonical upward axis of the image. This explicit spatial alignment reduces the cognitive burden of perspective transformation for the VLM.

A critical challenge in traditional BEV projection is the loss of the vertical (Z-axis) dimension. To implicitly preserve height relationships and relative occlusions, we compute a topological order between objects based on their Z-axis bounds, and then render the polygons $\mathcal{H}_k$ with a semi-transparent opacity and solid black borders. This visual design enhances the distinguishability of overlapping objects. For explicit height reasoning triggered by τ_{hgt}, we extract the bounding box Z-coordinates (z_{min}, z_{max}) of relevant instances to formulate a supplementary textual height descriptor $\mathcal{Z}_{hgt}$.

Micro-level Visual Detail Patches (VDP) While SSL establishes a robust spatial memory and serves as a reasoning medium for global spatial relationships, verifying attributes like material, state, or fine-grained shape necessitates high-fidelity visual inputs. Unlike previous methods that attempt to perceive spatial relations within local views, our VDP focuses *exclusively* on the occlusion-free visual appearance of the object itself.

For a selected instance o_k, we aim to select the optimal observation view v^* from the available RGB sequences $\mathcal{V}$. To ensure visual clarity and reduce computational overhead, we introduce a two-stage viewpoint selection mechanism. First, to filter out motion blur, we perform a sharpness-aware pre-filtering on the raw sequences. We divide the sequence into small temporal windows and retain only the frame with the highest sharpness—measured by the variance of

the Laplacian over the grayscale image—within each window. This step yields a high-quality, downsampled candidate view set $\hat{\mathcal{V}}$.

Subsequently, we evaluate the geometric exposure and image composition for each candidate view. The optimal viewpoint v^* is determined by maximizing a joint fitness function over the candidate set $\hat{\mathcal{V}}$:

$$v^* = \arg\max_{v \in \hat{\mathcal{V}}} \Big(\mathcal{S}_{geo}(v) \cdot \mathcal{S}_{comp}(v) \Big) \tag{1}$$

Geometry Exposure ($\mathcal{S}_{geo}$): We project the 3D point cloud P_k to the image plane of view v. A point $p \in P_k$ is deemed visible if the difference between its projected depth and the sensor depth is within a margin δ. Let $P_{vis}(v)$ denote the visible point set. $\mathcal{S}_{geo}$ integrates the basic visible ratio and the normal-weighted surface exposure:

$$\mathcal{S}_{geo}(v) = \alpha \frac{|P_{vis}(v)|}{|P_k|} + \beta \frac{1}{|P_k|} \sum_{p \in P_{vis}(v)} \max(0, \mathbf{n}_p \cdot \mathbf{d}_p) \tag{2}$$

where $\mathbf{n}_p$ is the surface normal, $\mathbf{d}_p$ is the viewing direction vector, and α, β are balancing coefficients. This encourages selecting views facing the object's broad surfaces.

Composition Truncation ($\mathcal{S}_{comp}$): A well-composed view should capture the object at a proper scale. We compute the area ratio r of the projected 2D bounding box relative to the image size. $\mathcal{S}_{comp}(v)$ employs a trapezoidal decay function to penalize extreme scales (too small or over-magnified), multiplied by a truncation penalty if the bounding box exceeds the image boundaries.

Once the optimal frame $I(v^*)$ is determined, the region of interest is cropped based on the projected 2D bounding box, with a slight expansion ratio to preserve essential local context. Guided by the parsed perception-demand flags, we selectively activate this intensive extraction process *only* for objects that are target candidates ($\tau_{tgt} = \text{True}$) or strictly require visual verification ($\tau_{vis} = \text{True}$). Finally, the crops of these objects are concatenated into a compact, unified image patch I_{VDP}. This approach provides low-noise, discriminative visual features while reducing token consumption, eliminating the necessity of resolving spatial relationships within complex local views.

3.4 Joint Spatial-Visual Reasoning

Equipped with the dual-granularity representation, we introduce a synergistic spatial-visual reasoning process. The query $\mathcal{Q}$, the macro-layout I_{SSL}, the micro-details I_{VDP}, along with supplementary textual descriptors (e.g., the height descriptors $\mathcal{Z}_{hgt}$), are simultaneously fed into the VLM.

As the preceding stages have explicitly decoupled spatial layout with visual appearance, the downstream reasoning process is greatly simplified. Consequently, rather than relying on complex heuristic rules or heavily engineered prompts, our joint reasoning can be executed through a straightforward instruction.

Specifically, we prompt the VLM to adopt a natural "Spatial-First, Visual-Second" workflow. The instruction explicitly guides the model to *"Use the first image to determine the spatial relationship, then use the second image to verify visual attributes."* To facilitate logical deduction, we simply request the VLM to generate a free-form rationale ("Analysis") prior to predicting the final target ID.

Driven by this concise instruction, the VLM naturally exhibits a zero-shot Chain-of-Thought (CoT) reasoning capability, executing a simple two-step reasoning pipeline:

1. **Spatial Deduction:** Analyzing the occlusion-free layout (I_{SSL}) to deduce spatial relation constraints and isolate a minimal set of candidate IDs.
2. **Visual Verification:** Inspecting the detail patches (I_{VDP}) to match fine-grained attributes of the specific candidates, and verify the target candidates, which is performed without the distraction of spatial crowding.

This streamlined reasoning process demonstrates that providing appropriately structured and normalized representations enables robust zero-shot localization, bypassing the need for complex, predefined reasoning templates.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our proposed framework on two widely adopted 3D visual grounding benchmarks: ScanRefer [3] and Nr3D [1]. Both datasets are built upon the ScanNet [5] dataset, which contains rich 3D scans of real-world indoor scenes.

- **ScanRefer** contains 51,583 language queries describing 11,046 objects. The queries are categorized into *Unique* (only one instance of the target category exists) and *Multiple* (distractors of the same category exist).
- **Nr3D** includes 41,503 queries. The queries are split into *Easy* (single distractor) and *Hard* (multiple distractors), as well as *View-dependent* and *View-independent*.

Evaluation Metrics. Following standard evaluation protocols, for ScanRefer, we measure grounding performance using Intersection over Union (IoU) and report accuracy at thresholds of 0.25 (Acc@0.25) and 0.5 (Acc@0.5). For Nr3D, the task is formulated as multiple-choice selection among given object proposals. We report the absolute classification accuracy.

Implementation Details. Unless otherwise specified, we employ gpt-4o-2024-08-06 [17] as our core engine for a fair comparison.

To reduce randomness and improve reasoning stability, we set the `temperature` to 0.1 and `top_p` to 0.3 across all prompts. For scene initialization, we follow prior zero-shot works [14, 29] by employing a pre-trained Mask3D [21] to obtain

instance proposals for ScanRefer. During the Visual Detail Patches (VDP) extraction, optimal 2D bounding boxes are expanded by a ratio of 1.5 to preserve local context. Additional details and prompts are provided in the Supplementary Material. The code is available on our project page [5].

4.2 Quantitative Results

Table 1: 3D visual grounding performance on the ScanRefer dataset. We group the methods by learning paradigms. Best zero-shot results are highlighted in **bold**.

Method	LLM / VLM	Unique		Multiple		Overall	
		@0.25	@0.5	@0.25	@0.5	@0.25	@0.5
Supervised Methods							
ScanRefer [3]	-	67.6	46.2	32.1	21.3	39.0	26.1
3DVG-Trans. [30]	-	81.9	60.6	39.3	28.4	47.6	34.7
BUTD-DETR [10]	-	84.2	66.3	46.6	35.1	52.2	39.8
ChatScene [8]	Vicuna-7B	89.6	82.5	47.8	42.9	55.5	50.2
Video-3D LLM [31]	LLaVA-Video 7B	88.0	78.3	50.9	45.3	58.1	51.7
GPT4Scene [19]	Qwen2-VL-7B	90.3	83.7	56.4	50.9	62.6	57.0
Zero-Shot Methods							
OpenScene [18]	CLIP	20.1	13.1	11.1	4.4	13.2	6.5
LLM-Grounder [24]	GPT-4 turbo	-	-	-	-	17.1	5.3
ZSVG3D [29]	GPT-4 turbo	63.8	58.4	27.7	24.6	36.4	32.7
SeeGround [14]	Qwen2-VL-72B	**75.7**	**68.9**	34.0	30.0	44.1	39.4
CSVG [28]	Mistral-Large-2407	68.8	61.2	38.4	27.3	49.6	39.8
Ours	GPT-4o	73.9	66.8	**45.6**	**39.7**	**51.1**	**45.0**

Comparison on ScanRefer Tab. 1 reports the results on ScanRefer. Our framework achieves the best zero-shot overall performance among the compared methods, with an Acc@0.5 of 45.0%. The improvement is mainly reflected in the *Multiple* subset, where the target category contains distractors and grounding requires relational disambiguation beyond category recognition. In this setting, our method obtains 39.7% Acc@0.5, improving over prior zero-shot methods by a clear margin. In contrast, on the *Unique* subset, where recognizing the target category is often sufficient, our performance is comparable to the strongest view-based baseline but not the highest. This result empirically validates our hypothesis that Semantic Spatial Layout (SSL) effectively resolves complex relation ambiguities that 2D view-based methods fail to capture.

[5] `https://github.com/VIPL-VSU/ALFD`

Table 2: 3D visual grounding performance on the Nr3D dataset. Methods are evaluated with or without ground-truth (GT) object classes. Best zero-shot results are in **bold**.

Method	LLM / VLM	Easy	Hard	Dep.	Indep.	Overall
Supervised Methods						
ReferIt3DNet [1]	-	43.6	27.9	32.5	37.1	35.6
3DVG-Trans. [30]	-	48.5	34.8	34.8	43.7	40.8
ViL3DRel [4]	-	70.2	57.4	62.0	64.5	64.4
3D-VisTA [33]	-	72.1	56.7	61.5	65.1	64.2
MiKASA [2]	-	69.7	59.4	65.4	64.0	64.4
SceneVerse [11]	-	72.5	57.8	56.9	67.9	64.9
Zero-Shot Methods (w/o GT object class)						
ZSVG3D [29]	GPT-4 turbo	46.5	31.7	36.8	40.0	39.0
SeeGround [14]	Qwen2-VL-72B	54.5	38.3	42.3	48.2	46.1
LaSP [16]	GPT-4o	60.7	45.3	49.2	54.7	52.9
Ours	GPT-4o	**70.9**	**55.9**	**59.2**	**65.2**	**63.2**
Zero-Shot Methods (w/ GT object class)						
CSVG [28]	Mistral-Large	67.1	51.3	53.0	62.5	59.2
Transcrib3D [6]	GPT-4	79.7	60.3	60.1	75.4	70.2
LaSP [16]	GPT-4o	76.3	59.6	61.6	71.0	67.8
Ours	GPT-4o	**82.9**	**66.9**	**69.0**	**77.6**	**74.7**

Comparison on Nr3D Tab. 2 summarizes the results on Nr3D under two evaluation settings. Without ground-truth object classes, our framework achieves 63.2% overall accuracy, outperforming existing zero-shot methods. The gains are particularly clear on the *Hard* and *View-dependent* subsets, where the language descriptions typically require distinguishing among multiple candidates or resolving viewpoint-sensitive spatial relations. These results suggest that representing the scene as a global semantic layout provides useful spatial evidence that is difficult to obtain from local camera views alone. When ground-truth object classes are provided, our framework further improves to 74.7% overall accuracy. This setting reduces the influence of perception and better isolates the effect of spatial-visual reasoning. The consistent improvement in both settings further indicates the benefit of our framework in modeling challenging spatial relationships.

Comparison on Subsets and Efficiency Recent agent-based methods [12, 23] adopt an alternative evaluation protocol proposed in VLM-Grounder [23], evaluating on sampled subsets of ScanRefer and Nr3D rather than full validation splits. For fair comparison, we evaluate our method on the identical subset. As shown in Tab. 3, our framework achieves competitive or better accuracy while requiring lower average inference time per query. This efficiency mainly comes from the compact dual-granularity representation: rather than processing long

Table 3: 3D visual grounding performance and efficiency on subsets of ScanRefer and Nr3D. We compare our framework against agent-based approaches. Inference time is measured in average seconds per query. Methods are evaluated without GT object classes on Nr3D. Best results are in **bold**.

Method	VLM	ScanRefer Subset	Nr3D Subset	Inf. Time ↓
		Overall Acc@0.5	Overall Acc	(s / query)
VLM-Grounder [23]	GPT-4o	33.5	48.0	50.3
SPAZER [12]	GPT-4o	48.8	63.8	23.5
Ours	GPT-4o	**49.2**	**66.0**	**10.9**

sequences of raw multi-view observations, the VLM receives a task-aware spatial layout and a small set of object-centric detail patches. As a result, the reasoning input is shorter and more directly aligned with the grounding decision.

4.3 Ablation Studies

Table 4: Ablation study on the architectural components. We evaluate the impact of Query-Guided Scene Filtering, Semantic Spatial Layout (SSL), and Visual Detail Patches (VDP) on the Nr3D subset. "Raw BEV" denotes directly rendering the unfiltered, unabstracted full scene from a top-down view with ID overlays.

Our Components			**Spatial Representation**	**Overall Acc (%)**
Filtering	SSL	VDP		
-	-	-	Raw BEV + ID (Full Scene)	48.8
✓	✓	-	Semantic Spatial Layout	60.8
✓	-	✓	None (Crops Only)	56.0
✓	✓	✓	Semantic Spatial Layout	**66.0**

To validate the effectiveness of our core designs, we conduct ablation studies on Nr3D subset without using ground-truth label names. We isolate the contribution of each core component in Table 4. As our baseline (Row 1), we directly render the entire unconstrained 3D scene from a top-down view (Raw BEV) with ID overlays but without any filtering or abstraction. This naive holistic view yields only 48.8% accuracy, as the VLM is overwhelmed by irrelevant visual clutter and complex physical occlusions.

When we introduce Query-Guided Filtering and abstract the scene into our Semantic Spatial Layout (SSL) *without* providing local patches (Row 2), the performance increases to 60.8% (+12.0%). This improvement suggests that removing background noise and converting the remaining objects into a clean semantic layout improves the VLM's spatial reasoning potential. Conversely, relying exclusively on Visual Detail Patches (VDP) after filtering (Row 3, 56.0%) improves

upon the baseline by eliminating noise but falls short of SSL. This is expected as isolated local crops do not explicitly encode room-level spatial context, failing to resolve complex, long-range dependencies. Finally, integrating both the macro-level SSL and micro-level VDP (Row 4), our full framework achieves the optimal 66.0%. This confirms that spatial relations and visual attributes are mutually complementary, and our dual-granularity decoupling is effective for 3D visual grounding.

4.4 Qualitative Results

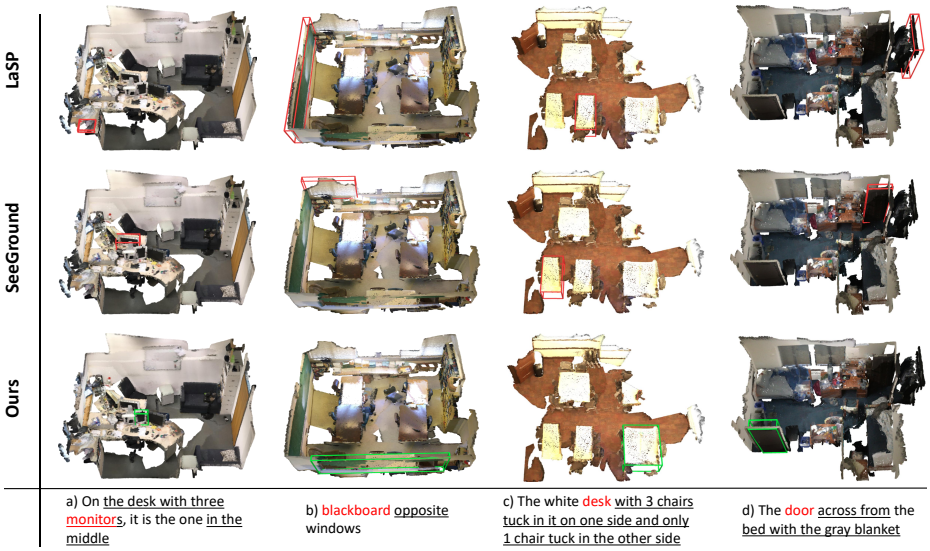

Fig. 3: Qualitative Results. Comparison of our framework with representative view-based and symbolic-based methods on challenging 3DVG queries involving long-range and compositional spatial reasoning.

Fig. 3 visualizes the grounding result of our method compared to state-of-the-art *symbol-based* (LaSP [16]) and *view-based* (SeeGround [14]) methods in some representative challenging scenarios. **1) Resolving Compositional Spatial Relations.** Cases (a) and (c) involve dense arrangements with numerical constraints (e.g., *"3 chairs tuck in..."*). View-based methods (SeeGround) can be unreliable because severe occlusions in 2D views make counting and relative-position reasoning difficult. Symbolic methods (LaSP) struggle to encode intricate spatial concepts like "tucked in" or "middle" into executable code due to lacking visual intuition. Conversely, the transparent Semantic Spatial Layout (SSL) explicitly exposes these asymmetrical distributions, allowing the VLM to count and resolve local topologies with less dependence on occlusion-prone views

or predefined relation rules. **2) Capturing Long-Range Spatial Dependencies.** Cases (b) and (d) demand macro-level reasoning (e.g., *"opposite windows"*). View-based methods suffer from spatial myopia, unable to simultaneously capture distant objects without severe perspective distortion. Symbolic methods possess global coordinates but often approximate topological terms (*e.g., "across from"*) with predefined distance heuristics. Ours effectively addresses this by using SSL as a global cognitive map where room-level spatial configurations are more directly exposed. Simultaneously, our Visual Detail Patches (VDP) supply high-fidelity object-centric crops to verify fine-grained attributes (e.g., *"gray blanket"*). This suggests how our dual-granularity decoupling helps reduce spatial ambiguity. Additional qualitative demonstrations of our framework in unconstrained, self-scanned real-world environments are provided in the Supplementary Material.

5 Conclusion

In this paper, we presented a dual-granularity framework for zero-shot 3D Visual Grounding. To overcome the spatial myopia and visual clutter, our framework aligns 3D scene abstraction with the human visual information-seeking principle. By proactively filtering the unconstrained scene and constructing a dual-granularity representation, we instantiate a holistic Semantic Spatial Layout as an explicit global cognitive medium, orthogonally complemented by Visual Detail Patches for fine-grained verification. This explicit decoupling successfully translates abstract spatial memory into an actionable, VLM-friendly format. Extensive experiments on benchmarks and real-world environments validate the effectiveness of our framework, especially on queries requiring long-range spatial dependencies and compositional spatial reasoning, demonstrating its potential to advance 3DVG.

Acknowledgements. This work is partially supported by Natural Science Foundation of China under contracts Nos. 62495082, 62461160331, and Beijing Municipal Natural Science Foundation Nos. L257009, L242025.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In: Proceedings of the European Conference on Computer Vision. pp. 422–440 (2020)
2. Chang, C.P., Wang, S., Pagani, A., Stricker, D.: MiKASA: Multi-key-anchor & scene-aware transformer for 3D visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14131–14140 (2024)
3. Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D object localization in RGB-D scans using natural language. In: Proceedings of the European Conference on Computer Vision. pp. 202–221 (2020)

4. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Language conditioned spatial relation reasoning for 3D object grounding. In: Advances in Neural Information Processing Systems. vol. 35, pp. 20522–20535 (2022)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5828–5839 (2017)
6. Fang, J., Tan, X., Lin, S., Vasiljevic, I., Guizilini, V., Mei, H., Ambrus, R., Shakhnarovich, G., Walter, M.R.: Transcrib3D: 3D referring expression resolution through large language models. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 9737–9744 (2024)
7. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., Gan, C., de Melo, C.M., Tenenbaum, J.B., Torralba, A., Shkurti, F., Paull, L.: ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In: IEEE International Conference on Robotics and Automation. pp. 5021–5028 (2024)
8. Huang, H., Chen, Y., Wang, Z., Huang, R., Xu, R., Wang, T., Liu, L., Cheng, X., Zhao, Y., Pang, J., Zhao, Z.: Chat-Scene: Bridging 3D scene and large language models with object identifiers. In: Advances in Neural Information Processing Systems. vol. 37, pp. 113991–114017 (2024)
9. Huang, P.H., Lee, H.H., Chen, H.T., Liu, T.L.: Text-guided graph neural networks for referring 3D instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1610–1618 (2021)
10. Jain, A., Gkanatsios, N., Mediratta, I., Fragkiadaki, K.: Bottom up top down detection transformers for language grounding in images and point clouds. In: Proceedings of the European Conference on Computer Vision. pp. 417–433 (2022)
11. Jia, B., Chen, Y., Yu, H., Wang, Y., Niu, X., Liu, T., Li, Q., Huang, S.: SceneVerse: Scaling 3D vision-language learning for grounded scene understanding. In: Proceedings of the European Conference on Computer Vision. pp. 289–310 (2024)
12. Jin, Z., Tu, R.C., Liao, J., Sun, W., Luo, X., Liu, S., Tao, D.: SPAZER: Spatial-semantic progressive reasoning agent for zero-shot 3D visual grounding. In: Advances in Neural Information Processing Systems. vol. 38, pp. 165549–165576 (2025)
13. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: LERF: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19729–19739 (2023)
14. Li, R., Li, S., Kong, L., Yang, X., Liang, J.: 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. pp. 3707–3717 (2025)
15. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 3D-SPS: Single-stage 3D visual grounding via referred point progressive selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16454–16463 (2022)
16. Mi, B., Wang, H., Wang, T., Chen, Y., Pang, J.: Language-to-space programming for training-free 3D visual grounding. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 3844–3864 (2025)
17. OpenAI: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
18. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: OpenScene: 3D scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 815–824 (2023)

19. Qi, Z., Zhang, Z., Fang, Y., Wang, J., Zhao, H.: GPT4Scene: Understand 3D scenes from videos with vision-language models. arXiv preprint arXiv:2501.01428 (2025)
20. Roh, J., Desingh, K., Farhadi, A., Fox, D.: LanguageRefer: Spatial-language model for 3D visual grounding. In: Proceedings of the 5th Conference on Robot Learning. pp. 1046–1056 (2022)
21. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3D: Mask transformer for 3D semantic instance segmentation. In: IEEE International Conference on Robotics and Automation. pp. 8216–8223 (2023)
22. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages. pp. 336–343 (1996)
23. Xu, R., Huang, Z., Wang, T., Chen, Y., Pang, J., Lin, D.: VLM-Grounder: A VLM agent for zero-shot 3D visual grounding. In: Proceedings of the 8th Conference on Robot Learning. pp. 3961–3985 (2025)
24. Yang, J., Chen, X., Qian, S., Madaan, N., Iyengar, M., Fouhey, D.F., Chai, J.: LLM-Grounder: Open-vocabulary 3D visual grounding with large language model as an agent. In: IEEE International Conference on Robotics and Automation. pp. 7694–7701 (2024)
25. Yang, J., Ding, R., Deng, W., Wang, Z., Qi, X.: RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19823–19832 (2024)
26. Yang, L., Yuan, C., Zhang, Z., Qi, Z., Xu, Y., Liu, W., Shan, Y., Li, B., Yang, W., Li, P., Wang, Y., Hu, W.: Exploiting contextual objects and relations for 3D visual grounding. In: Advances in Neural Information Processing Systems. vol. 36, pp. 49542–49554 (2023)
27. Yang, Z., Zhang, S., Wang, L., Luo, J.: SAT: 2D semantics assisted training for 3D visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1856–1866 (2021)
28. Yuan, Q., Li, K., Zhang, J.: Solving zero-shot 3D visual grounding as constraint satisfaction problems. arXiv preprint arXiv:2411.14594 (2024)
29. Yuan, Z., Ren, J., Feng, C.M., Zhao, H., Cui, S., Li, Z.: Visual programming for zero-shot open-vocabulary 3D visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20623–20633 (2024)
30. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2928–2937 (2021)
31. Zheng, D., Huang, S., Wang, L.: Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8995–9006 (2025)
32. Zhou, M., He, C., Wang, R., Chen, X.: OV3D-CG: Open-vocabulary 3D instance segmentation with contextual guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5305–5314 (2025)
33. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3D-VisTA: Pre-trained transformer for 3D vision and text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2911–2921 (2023)