OMG: Open-Modality Training-Free 3D Visual Grounding

Anonymous authors
Paper under double-blind review

Abstract

3D visual grounding (3DVG) localizes objects in 3D scenes from natural-language queries, supporting applications in robotics and augmented reality. Existing approaches typically couple scene representations and reasoning procedures to specific input modalities, limiting their applicability across sensing conditions. To address this limitation, we introduce OMG, an open-modality, training-free 3DVG framework that uses pretrained language and vision–language models to ground objects from point clouds, posed RGB-D observations, and unposed multi-view RGB images. OMG constructs a Modality-Unified Scene Interface that represents different inputs as semantic 3D instances linked to local evidence, enabling a shared grounding procedure. We introduce two complementary modules: Query-Conditioned Spatial Abstraction, which captures query-relevant spatial relations in a shared representation, and Adaptive Evidence Verification, which selectively incorporates instance-level visual and geometric cues to resolve candidate ambiguity. Experiments on ScanRefer and Nr3D demonstrate the effectiveness of OMG across the supported input settings. On ScanRefer, it achieves 51.3/46.2%, 53.1/47.8%, and 44.0/34.1% Acc0.25/0.5 with point-cloud, RGB-D, and unposed RGB inputs, respectively, and achieves 50.2% Top-1 accuracy on Nr3D. Compared with modality-specific zero-shot baselines on ScanRefer, OMG achieves the best results under point-cloud and unposed RGB inputs while remaining competitive under RGB-D input.

1 Introduction

3D visual grounding (3DVG) aims to localize a target object in a 3D scene from a natural-language description (Chen et al., 2020; Achlioptas et al., 2020), supporting language-guided interaction with physical environments. It requires reasoning about spatial relations and visual or geometric cues to distinguish candidate objects.

Whether trained with 3DVG supervision or designed for zero-shot inference, existing methods typically assume a fixed form of scene observation. Specifically, supervised point-cloud methods reason over explicit 3D geometry, while recent foundation-model approaches also build on particular inputs, such as point-cloud renderings or captured RGB views (Zhao et al., 2021; Li et al., 2025; Xu et al., 2024). However, in embodied settings, the available observations depend on the sensing setup: a robot may provide a LiDAR point cloud or posed RGB-D observations, whereas a mobile device or lightweight agent may have access only to RGB images. Figure 1 illustrates this gap: existing point-cloud- and RGB-D-based approaches follow modality-specific pipelines, while a shared grounding procedure should accommodate point clouds, posed RGB-D, and RGB images.

Recent advances in pretrained large language models (LLMs) and vision–language models (VLMs) have enabled training-free 3DVG, which performs grounding without task-specific parameter updates or fine-tuning. Existing approaches instantiate this paradigm through visual programming (Yuan et al., 2024; Mi et al., 2025), geometric constraints (Yuan et al., 2025), and reasoning over point-cloud renderings or captured views (Li et al., 2025; Xu et al., 2024). Training-free grounding is related to, but distinct from, zero-shot evaluation: zero-shot methods may still use models trained on base categories or auxiliary data, whereas training-free methods perform no task-specific optimization during grounding (Yuan et al., 2024; Li et al., 2025). By reducing reliance on 3DVG-specific annotations, this paradigm offers a promising route to transfer across unseen scenes and sensing conditions. This

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

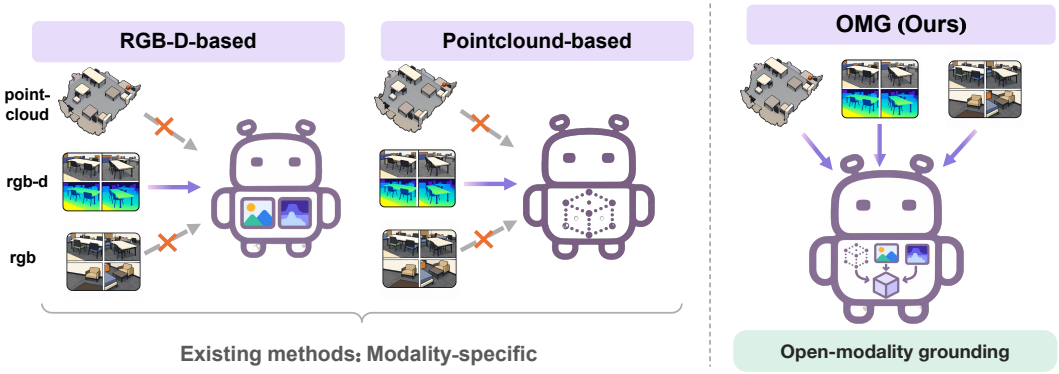


Figure 1: **Motivation for open-modality 3D visual grounding.** Existing point-cloud- and RGB-D-based methods use modality-specific pipelines, whereas OMG supports point-cloud, RGB-D, and RGB observations within one grounding framework.

motivates us to ask which information can support shared reasoning across input modalities. Object positions, extents, and target–reference configurations can be expressed in a common 3D representation, whereas the availability and quality of appearance and fine-grained geometric cues depend on the observation source. This distinction motivates a shared instance-level spatial representation, complemented by modality-specific evidence when needed.

We propose **OMG**, an open-modality, training-free framework for point clouds, posed RGB-D observations, and unposed multi-view RGB images. Specifically, for inputs from different modalities, we leverage corresponding perception front-ends to convert each observation into semantic 3D instances associated with local evidence, yielding a shared scene interface. This interface enables a unified grounding procedure across modalities without training or fine-tuning on 3DVG annotations. Building on this shared interface, we introduce Query-Conditioned Spatial Abstraction to encode query-relevant target–reference relations into a common spatial representation, and Adaptive Evidence Verification to selectively incorporate instance-level visual and geometric cues when ambiguity remains. Together, these components establish a shared reasoning mechanism across input modalities while introducing fine-grained evidence only when necessary.

We evaluate OMG on ScanRefer and Nr3D. On the full ScanRefer validation set, it achieves 51.3/46.2%, 53.1/47.8%, and 44.0/34.1% Acc0.25/0.5 with point-cloud, RGB-D, and unposed RGB inputs, respectively. Compared with modality-specific zero-shot methods, OMG achieves state-of-the-art performance under point-cloud and unposed RGB inputs, while remaining competitive in the RGB-D setting, demonstrating the effectiveness of unified open-modality grounding.

Our contributions are threefold:

- **Open-modality training-free 3DVG.** We introduce a shared semantic 3D instance interface that decouples modality-specific perception from grounding, enabling the same reasoning procedure across three input modalities.
- **Shared spatial reasoning with adaptive verification.** Query-Conditioned Spatial Abstraction preserves query-relevant target–reference configurations, while Adaptive Evidence Verification selectively incorporates local appearance and fine-grained geometric cues.
- **Evaluation across input modalities.** Experiments on ScanRefer and Nr3D demonstrate the effectiveness of the framework, with controlled ablations supporting the use of spatial abstraction and local evidence.

2 Related Work

Training-free 3D visual grounding. Early supervised 3D visual grounding methods learn language–geometry correspondence and contextual object relations from task-specific annotations (Chen et al., 2020; Zhao et al., 2021; Huang et al., 2022; Jain et al., 2022). Recent models further

strengthen multi-anchor and view-aware spatial reasoning, as in MiKASA and ViewSRD (Chang et al., 2024; Huang et al., 2025). In parallel, pretrained foundation models have enabled grounding without task-specific parameter updates. LLM-Grounder decomposes referring expressions and coordinates external grounding tools, ZSVG3D casts grounding as visual programming, while CSVG and LaSP externalize spatial reasoning through constraints or executable code (Yang et al., 2024; Yuan et al., 2024; 2025; Mi et al., 2025). In contrast, we encode query-relevant spatial structure as a compact visual abstraction for a frozen VLM and invoke local visual evidence only when needed. Here, *training-free* denotes the absence of task-specific parameter updates in the grounding pipeline.

Spatial representation and evidence organization. VLM-based grounding methods differ in how spatial context and object-level evidence are exposed during reasoning. SeeGround combines scene renderings with spatially enriched descriptions, while SPAZER progressively narrows candidates before incorporating relevant camera observations (Li et al., 2025; Jin et al., 2025). View-on-Graph instead organizes scene information as a multimodal graph for selective evidence retrieval (Liu et al., 2026), and recent agentic methods similarly invoke geometric or visual evidence on demand (Huynh et al., 2026; Zhang et al., 2026). Our approach differs in the representation itself: we construct a spatial abstraction that retains only target–reference configurations and coarse scene context, while local observations provide appearance and fine geometric cues. Candidate-adaptive routing further restricts expensive joint verification to ambiguous cases.

3D grounding from heterogeneous scene observations. Most existing grounding pipelines assume a particular form of scene input. VLM-Grounder operates on multi-view RGB images and recovers 3D predictions through multi-view projection, while Z3D extends image-based zero-shot grounding to settings with optional camera poses and depth (Xu et al., 2024; Drozdov et al., 2026). Beyond grounding, geometry foundation models such as DUSt3R and VGGT recover scene geometry and camera information directly from images, and SegVGGT further couples reconstruction with instance segmentation from unposed multi-view RGB observations (Wang et al., 2024; 2025b; Qu et al., 2026). We use pretrained perception front-ends to map point clouds, posed RGB-D, and unposed RGB into a common semantic instance interface, allowing the subsequent query interpretation, spatial abstraction, candidate selection, and verification procedure to remain shared across input modalities.

3 Method

3.1 Modality-Agnostic 3DVG

We study *modality-agnostic 3D visual grounding*, which does not assume a fixed input modality. Given a natural-language query q and a scene representation $\mathcal{S}_m$ in an arbitrary observation modality m, the goal is to localize the referred object:

$$\hat{B} = \mathcal{G}(\mathcal{S}_m, q), \tag{1}$$

where $\mathcal{G}$ denotes the complete grounding framework and $\hat{B} = (x, y, z, l, w, h)$ is an axis-aligned box in the scene coordinate frame. This setting is motivated by real-world embodied systems, where scene observations may come from different types of sensors and vary across platforms and environments. For example, robots may operate with LiDAR or RGB-D cameras, while mobile devices and lightweight agents may have access only to RGB images. A modality-agnostic grounding capability therefore enables a single system to interpret language and localize objects across diverse sensing conditions, making 3D grounding more broadly applicable to embodied navigation, robotic manipulation, augmented reality, and human–robot interaction. For this problem setting, we introduce OMG, an open-modality, training-free framework for 3D visual grounding.

Method Overview. As shown in Figure 2, our framework consists of three stages: (1) Modality-Unified Scene Interface, (2) Query-Conditioned Spatial Abstraction, and (3) Adaptive Evidence Verification. In the first stage, for any input modality, the corresponding perception front-end converts scene observations into a unified set of semantic 3D instances. These instances are further associated with modality-specific local evidence, forming a unified scene interface $\mathcal{I}_m$. In the Query-Conditioned Spatial Abstraction stage, given a referring query and the scene category vocabulary, an LLM performs role-aware category mapping to identify the target and reference categories and

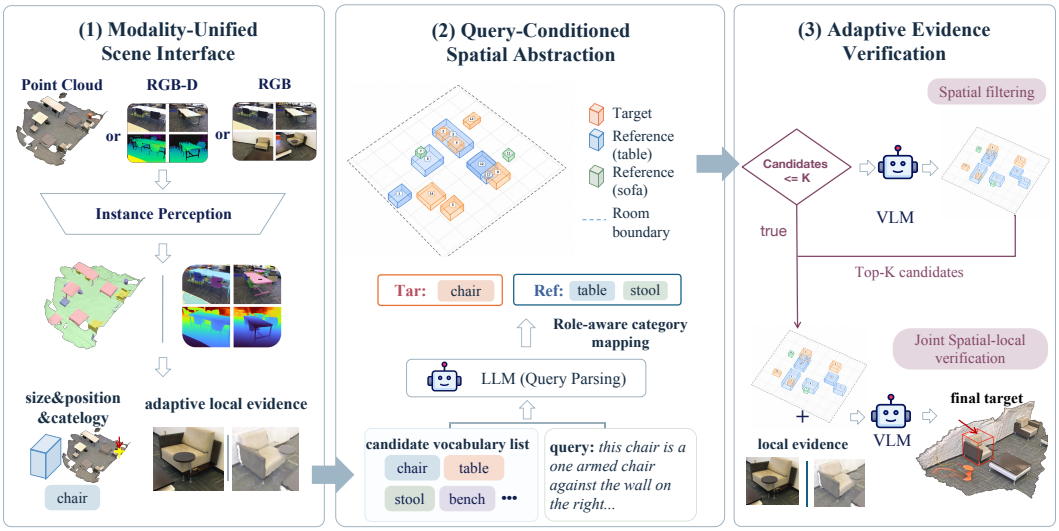


Figure 2: **Overview of OMG.** The framework unifies different input modalities at the instance level and models query-relevant spatial relations as a spatial abstraction, which serves as the core representation for subsequent reasoning and adaptive evidence verification.

extracts the query-relevant instances to construct a spatial abstraction. Finally, Adaptive Evidence Verification combines the spatial abstraction with the corresponding local evidence and, when necessary, performs joint verification through spatial filtering.

3.2 Modality-Unified Scene Interface

Different input modalities encode scene geometry, object semantics, and local visual cues in different forms. Accordingly, we handle modality-specific differences in the perception stage and expose a common object-level representation to subsequent grounding.

Point clouds are processed by a 3D instance segmentation model. Posed RGB-D observations are back-projected and fused in 3D before applying the same instance segmentation pipeline. For unposed RGB images, a multi-view perception model estimates geometry, camera parameters, and instance masks; these masks are lifted into a shared scene frame using the estimated geometry. After box fitting and duplicate suppression, the corresponding perception front-end for each modality m converts its scene observation $\mathcal{S}_m$ into a set $\mathcal{O}$ of semantic 3D instances:

$$\mathcal{O} = \mathcal{A}_m(\mathcal{S}_m) = \{o_i = (B_i, c_i)\}_{i=1}^{N}, \tag{2}$$

where $\mathcal{A}_m$ denotes the perception front-end for modality m, and B_i and c_i denote the 3D box and semantic category of instance i, respectively. The instance index i serves as its persistent ID. The full-scene extent B_{scene} is retained alongside $\mathcal{O}$. Therefore this set provides a common geometric and semantic representation for subsequent spatial reasoning.

Adaptive local evidence. While $\mathcal{O}$ captures the object-level structure required for spatial reasoning, distinguishing visually or geometrically similar candidates may require finer-grained cues from the corresponding input modality. For each instance o_i, we associate modality-specific local evidence $E_i^{(m)}$, yielding:

$$E_i^{(m)} = \Phi_m(\mathcal{S}_m, o_i), \quad \mathcal{I}_m = \{(B_i, c_i, E_i^{(m)})\}_{i=1}^{N}, \tag{3}$$

where Φ_m maps the scene observation $\mathcal{S}_m$ and instance o_i to the corresponding local evidence, and $\mathcal{I}_m$ denotes the resulting modality-unified scene interface. For inputs containing RGB observations, $E_i^{(m)}$ consists of image crops depicting the corresponding instance. For point-cloud-only input, $E_i^{(m)}$ is constructed from instance-centered virtual views of the point cloud to expose local geometric detail. Views are selected according to visibility, object coverage, and framing, independently of the

query text (Appendix B.1). This unified interface enables subsequent spatial reasoning and verification to operate consistently across input modalities, with modality-specific differences confined to perception and local evidence construction.

3.3 Query-Conditioned Spatial Abstraction

Existing VLM-based grounding methods typically reason over point-cloud renderings or captured RGB views, making their spatial representations dependent on a particular observation source. To support modality-independent spatial reasoning, we construct a query-conditioned spatial abstraction from the unified instance set $\mathcal{O}$, retaining only query-relevant instances. The construction consists of two steps: role-aware category mapping and spatial abstraction construction.

Role-aware category mapping. Literal matching between the query and scene categories may miss valid targets because of vocabulary mismatch, whereas overly permissive semantic matching can introduce distractor categories and unnecessarily enlarge the candidate set. We therefore perform category mapping over the category vocabulary present in $\mathcal{O}$. Let $\mathcal{C}$ denote the set of semantic categories in the scene. Given the query q and $\mathcal{C}$, an LLM jointly resolves the target and reference roles and maps them to the corresponding scene categories:

$$\mathcal{C} = \{c_i \mid o_i \in \mathcal{O}\}, \quad (\mathcal{C}_t, \mathcal{C}_r) = \mathcal{M}(q, \mathcal{C}), \tag{4}$$

where $\mathcal{C}_t$ and $\mathcal{C}_r$ denote the mapped target and reference categories, respectively, and $\mathcal{M}$ denotes the role-aware category mapping function. The mapping prioritizes exact or alias matches, followed by semantically compatible categories; when neither is available, a constrained fallback considers a plausible detector-confusion category for the same object concept. We then define the target instance set as $\mathcal{O}_t = \{o_i \in \mathcal{O} : c_i \in \mathcal{C}_t\}$. The reference set is defined analogously as $\mathcal{O}_r = \{o_i \in \mathcal{O} : c_i \in \mathcal{C}_r\}$.

Spatial abstraction construction. We construct a compact spatial abstraction that exposes their geometric relations without relying on the original input modality. Specifically, we render the selected instances as semi-transparent 3D cuboids:

$$X = \mathcal{R}(\mathcal{O}_t, \mathcal{O}_r), \tag{5}$$

where $\mathcal{R}$ denotes the abstraction renderer. Each cuboid preserves the center, spatial extent, and relative arrangement of its corresponding 3D instance, providing explicit cues for spatial relations such as distance, vertical ordering, relative scale, support, and containment. By retaining only query-relevant instances and removing texture and fine surface geometry, X emphasizes the spatial layout required while suppressing irrelevant visual detail.

The rendering further preserves the semantic roles established by role-aware category mapping. Target candidates are highlighted in orange, while reference instances are distinguished by category-specific colors, allowing the VLM to associate spatial configurations with their corresponding query roles. The abstraction is rendered from an elevated oblique viewpoint under orthographic projection, with the camera oriented toward the selected instances (Appendix B.2). This deterministic view provides a consistent visualization of the 3D configuration across input modalities; image-plane directions are not treated as semantic spatial directions during reasoning.

3.4 Adaptive Evidence Verification

Spatial abstraction captures the query-relevant relational structure of the scene, but spatial cues alone may be insufficient to distinguish candidates with similar configurations or differences primarily in appearance and fine-grained geometry. Applying fine-grained verification to all candidates, however, can substantially increase the visual context and require repeated access to local observations, particularly when the target set is large. We therefore adapt the verification procedure to the target-set size, using spatial abstraction to reduce ambiguity when necessary and introducing local evidence only for the instances retained for final verification.

Spatial filtering. Let $n = |\mathcal{O}_t|$, and let K denote the maximum number of candidates passed to joint verification. If $n = 0$, no valid target is returned; if $n = 1$, the unique candidate is directly

selected. When $1 < n \leq K$, spatial filtering is skipped and all target candidates proceed directly to joint verification. When $n > K$, the VLM uses the query q and query-conditioned spatial abstraction X to rank target candidates according to their spatial consistency with the query and retains the top K. This stage relies only on the abstract spatial representation, leaving appearance and fine-grained geometry to the subsequent verification step. Accordingly, for $n > 1$, the candidate set passed to joint verification is defined as

$$\mathcal{V} = \begin{cases} \mathcal{O}_t, & 1 < n \leq K, \\ \mathrm{TopK}_{\mathrm{VLM}}(q, X, \mathcal{O}_t, \mathcal{O}_r), & n > K. \end{cases} \tag{6}$$

where $\mathrm{TopK}_{\mathrm{VLM}}(\cdot)$ denotes the above spatial filtering operation.

Joint spatial–local verification. For the retained set $\mathcal{V}$, we update the query-conditioned spatial abstraction as $X_{\mathcal{V}} = \mathcal{R}(\mathcal{V}, \mathcal{O}_r)$ and pair each candidate $o_i \in \mathcal{V}$ with its corresponding local evidence $E_i^{(m)}$. The verification VLM then jointly reasons over the original query, the updated spatial abstraction, and these candidate–evidence pairs to select the referred instance and return its 3D box:

$$\hat{i} = F\Big(q, X_{\mathcal{V}}, \{(o_i, E_i^{(m)})\}_{o_i \in \mathcal{V}}\Big), \quad \hat{B} = B_{\hat{i}} \tag{7}$$

where F denotes the final VLM-based verification step. Spatial abstraction provides relational context, while local evidence supplies the appearance or fine-grained geometric cues needed for final disambiguation.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on two widely used 3D visual grounding benchmarks, ScanRefer (Chen et al., 2020) and Nr3D (Achlioptas et al., 2020). ScanRefer contains natural-language descriptions grounded to objects in ScanNet (Dai et al., 2017) scenes and evaluates direct 3D localization through predicted bounding boxes. We report results on the full validation set, comprising 9,508 referring expressions across 141 scenes, using Acc@0.25 and Acc@0.5 under the standard 3D IoU protocol. Nr3D, from ReferIt3D, focuses on fine-grained object disambiguation among candidate instances in a scene. Following the standard candidate-based protocol, we use the benchmark-provided object instances and report Top-1 grounding accuracy.

Implementation Details. We use modality-specific perception front-ends to convert different inputs into a unified set of semantic 3D object instances and bounding boxes. For point-cloud and RGB-D inputs, we use Mask3D (Schult et al., 2023) to obtain 3D instance masks and object proposals directly from scene geometry. For RGB-only inputs, SegVGGT (Qu et al., 2026) estimates instance-level geometry from multi-view images and reconstructs corresponding 3D object proposals. The resulting instances are then processed by the same grounding procedure across all input modalities. The grounding procedure requires no training or fine-tuning on 3DVG annotations; the perception front-ends are used with their pretrained weights. We use Qwen3-VL-32B-Instruct (Bai et al., 2025) with 4-bit AWQ quantization (Lin et al., 2024) for language and vision-language reasoning, served with vLLM (Kwon et al., 2023). Spatial filtering retains at most $K = 4$ candidate instances. The complete framework can run on a single NVIDIA RTX 3090 GPU. Additional inference-cost analysis and results with an alternative foundation model are provided in Appendix A.

4.2 Main Results under Different Input Modalities

Point-cloud only. Our method achieves **51.3/46.2** Acc@$0.25/0.5$ on ScanRefer (Table 1), improving over SeeGround (Li et al., 2025) by 7.2/6.8 points and UZ3DVG (Tan et al., 2026) by 5.9/5.1 points. On Nr3D (Table 2), it reaches **50.2%** Top-1 accuracy, compared with 46.1% for SeeGround (Li et al., 2025) and 46.5% for UZ3DVG (Tan et al., 2026). The consistent improvements across both benchmarks show that our unified grounding formulation remains effective without captured RGB observations or 3DVG-specific training.

RGB-D. On ScanRefer (Table 1), our method achieves **53.1/47.8** Acc@$0.25/0.5$, yielding competitive performance under the RGB-D setting, although it remains below the image-centric Z3D (Drozdov et al., 2026) (58.9/52.7). Unlike methods that directly exploit rich image context for spatial

Table 1: ScanRefer validation results (Acc, %). Acc@0.25 and Acc@0.5 denote localization accuracy at 3D IoU thresholds of 0.25 and 0.5, respectively. P, I, and RGB-D denote colored point clouds, captured RGB images, and RGB-D inputs with camera calibration, respectively. ZS denotes zero-shot grounding without ScanRefer-specific supervision, while Sup. denotes supervised grounding. UZ3DVG is reported with Mask3D refinement, and MC denotes MaskClustering. Bold values indicate the SOTA result within each comparison block, gray rows denote our method, and "–" indicates unavailable or unreported results.

Method	Input	Training	Proposals	Unique		Multiple		Overall	
				@0.25	@0.5	@0.25	@0.5	@0.25	@0.5
Supervised									
LIBA (Wang et al., 2025c)	P	Sup.	–	88.8	74.3	54.4	44.4	59.6	49.0
LLaVA-3D (Zhu et al., 2025)	RGB-D	Sup.	–	–	–	–	–	50.1	42.7
Video-3D LLM (Zheng et al., 2025b)	RGB-D	Sup.	–	86.6	77.0	50.9	45.0	57.9	51.2
ROSS3D (Wang et al., 2025a)	RGB-D	Sup.	–	87.2	77.4	54.8	48.9	61.1	54.4
GPT4Scene (Qi et al., 2026)	RGB-D	Sup.	–	**90.3**	**83.7**	**56.4**	**50.9**	**62.6**	**57.0**
VG-LLM (Zheng et al., 2025a)	I	Sup.	–	–	–	–	–	41.6	14.9
Zero-shot									
ZSVG3D (3D only) (Yuan et al., 2024)	P	ZS	Mask3D	57.1	49.4	25.9	23.3	33.1	29.3
CSVG (Yuan et al., 2025)	P	ZS	Mask3D	68.8	61.2	38.4	27.3	49.6	39.8
SeeGround (Li et al., 2025)	P	ZS	Mask3D	75.7	68.9	34.0	30.0	44.1	39.4
UZ3DVG (Tan et al., 2026)	P	ZS	Mask3D	73.6	68.2	40.5	36.3	45.4	41.1
OMG (Ours)	P	ZS	Mask3D	**82.1**	**75.1**	**41.5**	**37.0**	**51.3**	**46.2**
Zero-shot / posed RGB/D									
ZSVG3D (Yuan et al., 2024)	RGB-D	ZS	Mask3D	63.8	58.4	27.7	24.6	36.4	32.7
VoG-72B (Liu et al., 2026)	RGB-D	ZS	–	78.6	71.5	34.0	30.4	44.8	40.3
Z3D (Drozdov et al., 2026)	RGB-D	ZS	Mask3D	82.3	74.8	**51.5**	**45.7**	**58.9**	**52.7**
Z3D (Drozdov et al., 2026)	RGB-D	ZS	MC	73.9	64.0	47.8	40.3	54.2	46.0
OMG (Ours)	RGB-D	ZS	Mask3D	**83.2**	**76.0**	43.5	38.8	53.1	47.8
Zero-shot / unposed RGB									
DUSt3R (Wang et al., 2024) → SeeGround (Li et al., 2025)	I	ZS	–	35.2	17.5	13.9	5.2	19.0	8.2
DUSt3R (Wang et al., 2024) → Z3D (Drozdov et al., 2026)	I	ZS	MC	42.7	21.9	27.5	10.1	31.2	12.9
OMG (Ours)	I	ZS	SegVGGT	**73.1**	**59.3**	**34.7**	**26.1**	**44.00**	**34.1**

Table 2: Performance on the Nr3D validation set (Acc, %). We compare point-cloud-only methods (P) under the standard object-candidate protocol. Queries are categorized as Easy/Hard according to the number of distractors and as View-Dependent/View-Independent according to whether grounding requires a specific viewpoint.

Method	Venue	Input	Zero-shot	Easy	Hard	View-Dep.	View-Indep.	Overall
Non-zero-shot								
Vigor (Wu et al., 2025)	WACV'25	P	No	66.6	53.1	59.2	59.9	59.7
MiKASA (Chang et al., 2024)	CVPR'24	P	No	69.7	59.4	65.4	64.0	64.4
ViewSRD (Huang et al., 2025)	ICCV'25	P	No	**75.3**	**64.8**	**68.6**	**70.6**	**69.9**
Zero-shot								
ZSVG3D (3D only) (Yuan et al., 2024)	CVPR'24	P	Yes	45.9	27.9	34.9	38.4	36.7
SeeGround (Li et al., 2025)	CVPR'25	P	Yes	54.5	38.3	42.3	48.2	46.1
UZ3DVG (Tan et al., 2026)	CVPR'26	P	Yes	52.2	40.8	39.3	50.1	46.5
OMG (Ours)	–	P	Yes	**56.7**	**43.9**	**41.7**	**54.4**	**50.2**

reasoning, our approach performs spatial reasoning over a unified 3D representation and uses RGB observations primarily for appearance verification. This design allows the same grounding formulation to be applied consistently across modalities while retaining competitive performance under RGB-D input.

RGB only. With abstract 3D geometry recovered from unposed multi-view images, our method achieves **44.0/34.1** Acc@$0.25/0.5$ on ScanRefer (Table 1), outperforming DUSt3R (Wang et al., 2024)→Z3D (Drozdov et al., 2026) by 12.8/21.2 points. It also surpasses the supervised VG-LLM (Zheng et al., 2025a), which obtains 41.6/14.9. These results show that, without reconstructing a complete scene point cloud or requiring task-specific training, an abstract 3D geometric representation can still provide an effective structural basis for explicit spatial reasoning, while maintaining strong localization performance under the stricter IoU threshold.

4.3 Ablation Studies

4.3.1 Overall Component Ablation

We evaluate all configurations on the full ScanRefer validation set under the point-cloud-only setting. Table 3 ablates three components: role-aware category mapping, spatial abstraction, and joint spatial-local verification. When role-aware category mapping is removed, we replace it with stem-based and fuzzy string matching; when spatial abstraction is removed, we represent each candidate using textual descriptions of its 3D position and size. Without joint spatial-local verification, we directly take the Top-1 result after spatial filtering, without subsequent joint verification. Thus, each removed component is replaced by its corresponding baseline operation.

Table 3: Component ablation of role-aware category mapping, spatial abstraction, and joint spatial-local verification. ✓ and × indicate enabled and disabled components, respectively. Acc@0.25 and Acc@0.50 are reported in percent. The configurations include the baseline, incremental additions, and leave-one-component-out variants.

ID	Role-aware Mapping	Spatial Abstraction	Joint Verification	Acc@0.25	Acc@0.50
(a)	×	×	×	39.95	36.15
(b)	✓	×	×	44.55	39.91
(d)	✓	✓	×	46.97	41.92
(f)	✓	×	✓	50.63	45.58
(g)	×	✓	✓	45.82	41.63
(h)	✓	✓	✓	**51.27**	**46.18**

The full model improves Acc@0.25/0.50 over the baseline by 11.32/10.03 points, and removing any component reduces accuracy. Category mapping and joint verification have the largest leave-one-out effects: removing them costs 5.45/4.55 and 4.30/4.26 points, respectively. This supports their complementary roles in retaining the target during candidate construction and choosing among the retained candidates. Spatial abstraction has a smaller effect on final accuracy, but its gain is larger before verification ((d) vs. (b)) than after it ((h) vs. (f)). This pattern suggests that its main contribution lies in the first-stage candidate selection, with the subsequent verification reducing the effect of representation differences.

4.3.2 Ablation on Category Mapping

To evaluate the effectiveness of hierarchical semantic mapping, we compare three query-to-Mask3D category mapping strategies on all 9,508 ScanRefer validation queries: stem-based and fuzzy string matching, CLIP (Radford et al., 2021) Top-1 semantic matching, and our hierarchical semantic mapping. We construct candidate sets from the mapped Mask3D categories and report Recall@0.25/0.5 according to whether any candidate overlaps the ground-truth box at the corresponding IoU threshold; failed mappings yield empty candidate sets.

Table 4: Ablation on query-to-category mapping. ↑ indicates that higher values are better (Recall), while ↓ indicates that lower values are better (Empty Rate and average candidate count).

Category Mapping	Recall@0.25 ↑	Recall@0.5 ↑	Empty Rate ↓	Avg. # Cand. ↓
LLM Target + fuzzy and stem match	78.05	71.49	10.46	8.14
LLM Target + CLIP Top-1	81.09	73.83	**0.06**	**6.77**
Hierarchical Semantic Mapping	**85.40**	**77.83**	1.18	6.96

Hierarchical mapping achieves the highest Recall@0.25/0.5 (**85.40%/77.83%**). It outperforms lexical matching despite producing fewer candidates on average, and it improves recall over CLIP Top-1 with a similar candidate count. This suggests the gain comes from selecting more relevant categories rather than expanding the candidate pool. CLIP Top-1 has the lowest empty rate, but fewer of its candidate sets contain a box that meets the IoU threshold; avoiding empty mappings alone does not ensure target coverage.

4.3.3 Ablation on Spatial Representation

To evaluate the effectiveness of spatial abstraction for first-stage candidate selection, we compare bbox text, spatial abstraction, and point-cloud rendering while keeping the candidate pool, VLM backbone, and Top-K setting fixed. Both ScanRefer and Nr3D are evaluated on queries with more than four candidates, comprising 3,806 and 1,042 queries, respectively.

Table 5: Ablation of spatial representations on ScanRefer and Nr3D subsets.

Spatial Representation	ScanRefer				Nr3D	
	Top-1@0.25	Top-4@0.25	Top-1@0.50	Top-4@0.50	Top-1 Accuracy	Top-4 Accuracy
BBox Text	22.18	53.55	19.47	48.21	39.35	83.78
Spatial Abstraction	25.20	57.07	22.33	**52.00**	43.95	**86.47**
Point-cloud Rendering	**26.17**	**57.23**	**22.94**	51.89	**44.05**	84.84

Spatial abstraction consistently improves over bbox text and achieves performance comparable to point-cloud rendering. The slight Top-1 advantage of point-cloud rendering may stem from the additional fine-grained appearance and geometric details it preserves, while our Spatial Abstraction effectively captures query-relevant spatial information with a compact representation.

4.3.4 Ablation on Adaptive Evidence Verification

We further examine whether the fine-grained information omitted by spatial abstraction can be recovered through the proposed **Adaptive Evidence Verification**. To this end, we provide spatial abstraction and point-cloud rendering with the same local visual evidence and compare their final grounding performance.

Table 6: Ablation of Adaptive Evidence Verification on ScanRefer.

Spatial Representation	First-stage Top-1@0.25	First-stage Top-4@0.25	Final Acc@0.25	Final Acc@0.50
Spatial Abstraction	46.97	71.96	51.27	46.18
Point-cloud Rendering	**48.60**	**72.13**	**51.63**	**46.34**

Point-cloud rendering shows a slight advantage in first-stage candidate selection, while the two representations achieve nearly identical Top-4 recall. After introducing the same local evidence through Adaptive Evidence Verification, the final gaps are reduced to only 0.36 and 0.16 percentage points at Acc@0.25 and Acc@0.50, respectively. These results show that the fine-grained information omitted by spatial abstraction can be effectively complemented by local evidence during verification. They also show that our proposed spatial abstraction representation achieves nearly the same final grounding accuracy as point-cloud rendering.

5 Conclusion

We presented OMG, an open-modality, training-free framework for 3D visual grounding across point clouds, posed RGB-D observations, and unposed multi-view RGB images. OMG decouples modality-specific perception from shared grounding reasoning through a semantic 3D instance interface, allowing scene observations from different input modalities to follow a common grounding procedure. Query-conditioned spatial abstraction retains the target–reference structure needed for spatial reasoning, while candidate-adaptive evidence verification selectively introduces local appearance and geometric cues when candidate ambiguity remains. Experiments on ScanRefer and Nr3D demonstrate the effectiveness of this design, while controlled ablations show that the compact spatial abstraction preserves candidate coverage and achieves final accuracy comparable to point-cloud rendering when paired with the same local evidence. Overall, our results suggest that separating modality-specific scene perception from shared spatial reasoning, together with selective fine-grained evidence verification, provides a practical basis for 3D visual grounding across different input modalities.

AI Use Statement

Generative AI tools were used to aid and polish the manuscript writing, including suggesting edits to grammar, clarity, and phrasing. The authors reviewed the suggested edits and verified all citations, technical claims, numerical results, and final wording.

Reproducibility Statement

The paper describes the method, model settings, datasets, evaluation metrics, and experimental setup. The method requires no task-specific parameter updates. Code and configuration files will be provided with the final version to support reproduction.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123460409.pdf.

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

Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. MiKASA: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *CVPR*, pp. 14131–14140, 2024. URL https://birdy666.github.io/projects/mikasa/.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *European Conference on Computer Vision*, 2020. URL https://daveredrum.github.io/ScanRefer/.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

Nikita Drozdov, Andrey Lemeshko, Nikita Gavrilov, Anton Konushin, Danila Rukhovich, and Maksim Kolodiazhnyi. Z3D: Zero-shot 3D visual grounding from images. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 147–154, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-short.13. URL https://aclanthology.org/2026.acl-short.13/.

Ronggang Huang, Haoxin Yang, Yan Cai, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. ViewSRD: 3d visual grounding via structured multi-view decomposition. In *ICCV*, pp. 9726–9736, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Huang_ViewSRD_3D_Visual_Grounding_via_Structured_Multi-View_Decomposition_ICCV_2025_paper.pdf.

Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15524–15533, 2022. doi: 10.1109/CVPR52688.2022.01508. URL https://openaccess.thecvf.com/content/CVPR2022/html/Huang_Multi-View_Transformer_for_3D_Visual_Grounding_CVPR_2022_paper.html.

Cuong Huynh, Maxim Popov, Denis Gridusov, and Sergey Kolyubin. AgentGrounder: Zero-shot 3d visual pointcloud grounding using multimodal language models. *arXiv preprint arXiv:2605.25901*, 2026. URL https://arxiv.org/abs/2605.25901.

Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pp. 417–433, 2022. doi: 10.1007/978-3-031-20059-5_24. URL https://link.springer.com/chapter/10.1007/978-3-031-20059-5_24.

Zhao Jin, Rong-Cheng Tu, Jingyi Liao, Wenhao Sun, Xiao Luo, Shunyu Liu, and Dacheng Tao. SPAZER: Spatial-semantic progressive reasoning agent for zero-shot 3d visual grounding. In *NeurIPS*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f1ea808562a3cd5a2a0ea94f3dee675a-Paper-Conference.pdf.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. 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 (CVPR)*, pp. 3707–3717, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_SeeGround_See_and_Ground_for_Zero-Shot_Open-Vocabulary_3D_Visual_Grounding_CVPR_2025_paper.html.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Jiawen Lin, Shiran Bian, Yihang Zhu, Wenbin Tan, Yachao Zhang, Yuan Xie, and Yanyun Qu. Seqvlm: Proposal-guided multi-view sequences reasoning via vlm for zero-shot 3d visual grounding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3094–3103, 2025. doi: 10.1145/3746027.3754985.

Yuanyuan Liu, Haiyang Mei, Dongyang Zhan, Jiayue Zhao, Dongsheng Zhou, Bo Dong, and Xin Yang. View-on-graph: Zero-shot 3d visual grounding via vision-language reasoning on scene graphs. In *AAAI*, pp. 7386–7394, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/37677.

Boyu Mi, Hanqing Wang, Tai Wang, Yilun Chen, and Jiangmiao Pang. Language-to-space programming for training-free 3d visual grounding. In *EMNLP*, pp. 3844–3864, 2025. URL https://aclanthology.org/2025.emnlp-main.191/.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. GPT4Scene: Understand 3d scenes from videos with vision-language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html.

Jinyuan Qu, Hongyang Li, and Lei Zhang. Segvggt: Joint 3d reconstruction and instance segmentation from multi-view images. *arXiv preprint arXiv:2603.19926*, 2026. URL https://arxiv.org/abs/2603.19926.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8216–8223, 2023. doi: 10.1109/ICRA48891.2023.10160590.

Wenbin Tan, Jiawen Lin, Yuan Xie, Yachao Zhang, and Yanyun Qu. Uz3dvg: Unaided zero-shot 3d visual grounding with generated language conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9547–9557, June 2026.

Haibo Wang, Zihao Lin, Zhiyang Xu, and Lifu Huang. Think, act, build: An agentic framework with vision language models for zero-shot 3d visual grounding. *arXiv preprint arXiv:2604.00528*, 2026. URL https://arxiv.org/abs/2604.00528.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3D: Reconstructive visual instruction tuning with 3d-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9275–9286, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_Ross3D_Reconstructive_Visual_Instruction_Tuning_with_3D-Awareness_ICCV_2025_paper.html.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, June 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

Yuan Wang, Ya-Li Li, Eastman Z. Y. Wu, and Shengjin Wang. LIBA: Language instructed multi-granularity bridge assistant for 3d visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8114–8122, 2025c. doi: 10.1609/aaai.v39i8.32875. URL https://ojs.aaai.org/index.php/AAAI/article/view/32875.

Tung-Yu Wu, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Data-efficient 3d visual grounding via order-aware referring. In *WACV*, 2025. URL https://openaccess.thecvf.com/content/WACV2025/papers/Wu_Data-Efficient_3D_Visual_Grounding_via_Order-Aware_Referring_WACV_2025_paper.pdf.

Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. In *Conference on Robot Learning (CoRL)*, 2024. URL https://arxiv.org/abs/2410.13860.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F. Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7694–7701, 2024. doi: 10.1109/ICRA57147.2024.10610443. URL https://ieeexplore.ieee.org/document/10610443/.

Qihao Yuan, Kailai Li, and Jiaming Zhang. Solving zero-shot 3d visual grounding as constraint satisfaction problems. In *BMVC*, 2025. URL https://bmvc2025.bmva.org/proceedings/644/.

Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *CVPR*, pp. 20623–20633, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yuan_Visual_Programming_for_Zero-shot_Open-Vocabulary_3D_Visual_Grounding_CVPR_2024_paper.html.

Jiaxi Zhang, Yunheng Wang, Wei Lu, Taowen Wang, Shuning Zhang, Yixiao Feng, Junzhe Xu, Shunwang Sun, Weisheng Xu, Yuetong Fang, and Renjing Xu. Uniground: Universal 3d visual grounding via training-free scene parsing. *arXiv preprint arXiv:2603.08131*, 2026. URL https://arxiv.org/abs/2603.08131.

Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2928–2937, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_3DVG-Transformer_Relation_Modeling_for_Visual_Grounding_on_Point_Clouds_ICCV_2021_paper.html.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *Advances in Neural Information Processing Systems*, volume 38, pp. 23356–23382, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1d8dc55c1f6cf124af840ce1d92d1896-Abstract-Conference.html.

Duo Zheng, Shijia Huang, and Liwei Wang. 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 (CVPR)*, pp. 8995–9006, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zheng_Video-3D_LLM_Learning_Position-Aware_Video_Representation_for_3D_Scene_Understanding_CVPR_2025_paper.html.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering lmms with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4295–4305, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_LLaVA-3D_A_Simple_yet_Effective_Pathway_to_Empowering_LMMs_with_ICCV_2025_paper.html.

A Inference Efficiency

We analyze inference efficiency on a ScanRefer subset released with TAB-Agent (Wang et al., 2026), focusing on the number of model calls per query and Peak Context. Peak Context denotes the maximum input-plus-output token count of a single model call. These two measures reflect interaction cost and the computational and memory burden during inference, respectively. We compare our method with SeqVLM (Lin et al., 2025), which performs 3D visual grounding through multiple rounds of interaction with a VLM using real RGB images.

Table 7: Inference cost and localization accuracy on a ScanRefer subset (Wang et al., 2026).

Method	Model	Avg. calls ↓	Max. calls ↓	Peak Context ↓	Acc@0.25 ↑	Acc@0.50 ↑
OMG	Qwen3-VL 32B AWQ	2.168	**3**	**5,435**	53.2%	47.2%
OMG	Doubao Seed 2.1 Pro	**2.144**	**3**	6,552	**58.8**%	**53.2**%
SeqVLM (Lin et al., 2025)	Qwen3-VL 32B AWQ	2.740	14	10,055	50.0%	45.2%

With Qwen3-VL 32B AWQ, our method reduces the average number of model calls from 2.740 for SeqVLM to 2.168, the maximum number of calls per query from 14 to 3, and the Peak Context from 10,055 to 5,435 tokens. Overall, fewer model interactions and a more compact context substantially reduce inference and memory overhead. The model and inference procedure used in this work can run on a single RTX 3090 GPU with 24 GB of memory, demonstrating practical deployability under limited computational resources.

We further conduct a complementary experiment with the closed-source Doubao Seed 2.1 Pro model. Our method requires only 2.144 model calls per query on average and achieves 58.8/53.2 Acc@0.25/0.50. These results suggest that the proposed reasoning framework is not tied to a specific foundation model and shows further potential when paired with more capable general-purpose models.

B Additional Details

B.1 Local-Evidence Construction Details

RGB observations. We project candidate instance points into captured RGB-D frames using the camera intrinsics and poses, accounting for scene-coordinate alignment. Depth-consistency checks remove invalid and occluded projections. We fit a 2D bounding box to the remaining projections and discard views with insufficient projected support or small box dimensions. During preprocessing, we rank valid views by projected bounding-box area and retain a limited set of observations. At inference, we select the saved crop with the largest pixel area. This selection does not use an additional learned sharpness or visibility score.

Point-cloud-only input. We construct virtual views independently of the captured camera trajectory. For each candidate, we generate 360 camera configurations looking at its box center by combining 24 azimuths, five height offsets, and three distance scales. We prefer configurations whose projected box is sufficiently contained within the image and occupies a suitable area; if none satisfy these conditions, we retain the original camera-configuration pool. We shortlist up to 36 configurations using azimuth and height bins to encourage viewpoint diversity.

For each shortlisted view, we compare projected instance-point depths with a low-resolution scene-mesh depth rendering and compute

$$S_{\text{view}} = 0.68S_{\text{vis}} + 0.22S_{\text{cov}} + 0.10S_{\text{frame}}. \tag{8}$$

Here, S_{vis} is the fraction of unique projected target pixels supported by depth-consistent instance points, S_{cov} is the target's projected bounding-box coverage, clipped at 0.45 and normalized by 0.45, and S_{frame} is the fraction of instance points lying within the view. Coverage is computed from depth-consistent projections when sufficient support exists, and otherwise from all in-frame projections.

We select the highest-scoring view among those satisfying $S_{\text{frame}} \geq 0.85$; if this subset is empty, we select the highest-scoring evaluated view. We render the full colored scene mesh from the selected camera and overlay a red 2D rectangle enclosing the projected candidate box. If valid instance points are unavailable, camera selection falls back to projected-box framing quality.

B.2 Camera Construction for the Query-Conditioned Spatial Abstraction

We use a deterministic, query-conditioned oblique camera with orthographic projection. Let p_i denote the center of box B_i, and let I_t contain the target candidates displayed at the current stage. For a non-empty target set,

$$p_t = \frac{1}{|I_t|} \sum_{i \in I_t} p_i. \tag{9}$$

Let p_s be the center of the complete scene bounds. We lift this center by $\Delta h = 1.5$ along the vertical axis and construct the camera as

$$\widetilde{p}_s = p_s + (0, 0, \Delta h), \qquad p_{\text{cam}} = \widetilde{p}_s + (\widetilde{p}_s - p_t). \tag{10}$$

The camera looks toward p_t, using the scene's vertical axis to define its orientation. Its orthographic scale is derived from the complete scene bounds. This produces a shared view for the displayed target candidates rather than a separate spatial-abstraction camera for each candidate. The viewing direction depends on the candidate group and is not assumed to coincide with the viewpoint implied by the query; this distinction is explicitly stated in the reasoning prompt.

B.3 Role-First Category Alignment Prompt

We present the role-first category alignment prompt and its JSON output schema in Fig. 3. The prompt image includes the input placeholder. The `role_check` field is a model-generated self-check within the same call, not an independently verified guarantee.

Resolve object roles in a 3D visual grounding query and align them with Mask3D categories.

Complete the reasoning in this order:

1. Resolve the grammatical referent before considering detector categories.
2. Reduce the target phrase to its physical head noun by removing color, material, shape, size, and state modifiers.
3. Identify only concrete reference objects that participate in a relation with the target.
4. Map the fixed target and reference roles to available_mask3d_categories.
5. Confirm that no reference object replaced the target because its category was difficult to match.

Choose detector categories only from available_mask3d_categories. Do not invent categories.

Return strict JSON only, using this schema:

```
{
"target": {
"query_category": "complete target noun phrase from the caption",
"head_noun": "unmodified physical target category",
"exact_or_alias": ["..."],
"semantic_compatible": ["..."],
"possible_misclassified": ["..."]
},
"anchors": [
{
"query_category": "reference-object phrase from the caption",
"head_noun": "unmodified physical reference category",
"exact_or_alias": ["..."],
"semantic_compatible": ["..."],
"possible_misclassified": ["..."]
}
],
"role_check": {
"target_is_requested_object": true,
"target_was_not_replaced_by_anchor": true,
"brief_reason": "short grammatical role check"
}
}
```

Target-role rules:

- The target is the object the full description asks the listener to locate. It is not necessarily the last, most frequent, or most visually distinctive noun.
- Resolve "it," "this," "that," and omitted subjects to their antecedents before assigning roles.
- Objects introduced by "there is ... on it," "with ...," "next to ...," or "behind it" usually serve as references or context, not as replacements for the antecedent target.
- Example: "the bed is on the left; there is a laptop on it" — bed is the target; laptop is a reference.
- Example: "a brown couch with a bag on its cushion" — couch is the target; bag is a reference.
- Spatial words, room regions, appearance attributes, and target parts are not target categories.

Category-mapping rules:

- Map the target only after its grammatical role and head_noun are fixed.
- Strip modifiers before matching: "cylindrical table" maps to table; "black mattress" maps to mattress when those categories are available.
- Prefer one precise target detector category. Use at most two only for genuine aliases or equally plausible names for the same physical object.
- Use exact_or_alias for exact names, singular/plural forms, and common aliases such as couch/sofa or tv/television.
- Use semantic_compatible only for ordinary category variation such as desk/table or picture/painting/photo.
- Use possible_misclassified only when no exact or semantic category can represent the same target object. Never place a spatially related anchor category there.
- If exact_or_alias and semantic_compatible are both empty, choose exactly one best available detector-confusion category under possible_misclassified whenever a physical category could plausibly represent the target. Consider coarse shape, size, and typical room placement. This is a required single fallback, not permission to list multiple speculative categories.
- Examples when the true detector label is absent: mattress may be confused with bed/bench; a small cylindrical table with nightstand/dresser; cabinet with shelf.
- Leave the target category union empty only when no available physical category could plausibly represent the referred object.
- Keep distinct reference roles as separate anchor entries. Return an empty anchors list when no concrete reference is needed.
- Copy every returned detector category exactly from available_mask3d_categories.

Input:

{input_json}

Figure 3: Role-first category alignment prompt and its JSON output schema.