Zero-Shot 3D Visual Grounding from Vision-Language Models

Anonymous Submission to 3D-LLM/VLA Workshop @ CVPR 2025

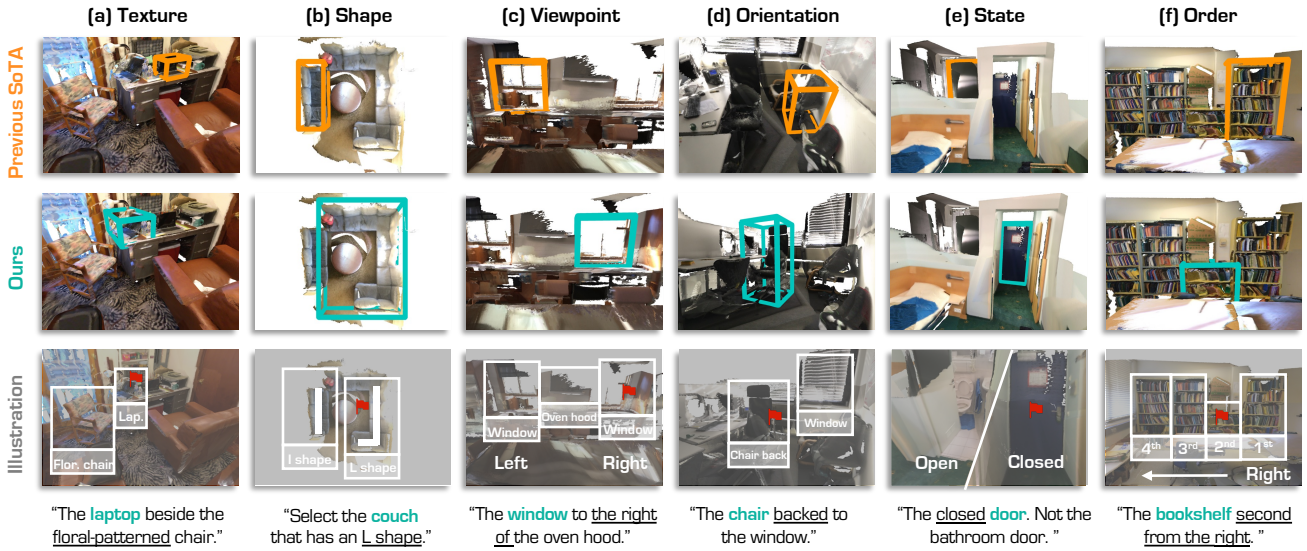


Figure 1. Effectiveness of ***SeeGround***: Different from previous SoTA, our method associates 2D visual cues – ***color***, ***texture***, ***viewpoint***, ***spatial position***, ***orientation***, and ***state*** – with 3D spatial text description to achieve precise scene understanding. Specifically, our method: (a) identifies the floral chair by recognizing unique color and texture cues; (b) recognizes the couch by interpreting geometric shape; (c) determines the right window by interpreting spatial relationships and perspective; (d) identifies the chair by discerning directional alignment; (e) detects the closed door by visually interpreting its state; and (f) selects the bookshelf by understanding relative positioning.

Abstract

*3D Visual Grounding (3DVG) aims to localize objects in 3D scenes based on textual descriptions, enabling applications in augmented reality and robotics. Traditional approaches rely on annotated 3D datasets and predefined object categories, limiting scalability. We introduce **SeeGround**, a zero-shot 3DVG framework leveraging 2D Vision-Language Models (VLMs). We bridge 3D scenes and 2D-VLM inputs via a hybrid representation of query-aligned rendered images and spatially enriched text. It introduces two key modules: the Perspective Adaptation Module for dynamic viewpoint selection, and the Fusion Alignment Module for aligning visual and spatial features to enhance localization. Extensive experiments on ScanRefer and Nr3D demonstrate that our approach outperforms existing zero-shot methods by large margins. Notably, we exceed weakly supervised methods and rival some fully supervised ones, outperforming previous SOTA by 7.7% on ScanRefer and 7.1% on Nr3D, showcasing its effectiveness in complex 3DVG tasks. The code will be made publicly available.*

1. Introduction

3D Visual Grounding (3DVG) aims to localize target objects in 3D scenes based on textual descriptions, supporting applications such as augmented reality [6, 40–42, 45, 60], vision-language navigation [11, 14, 21], and robotic perception [4, 7, 18, 28–34, 55, 76, 77]. Solving 3DVG requires both language understanding and spatial reasoning in complex environments. Most prior works train task-specific models [6, 23, 50, 61, 70, 73, 74] on small-scale datasets, limiting generalization. Scaling to real-world scenes is costly [3, 12, 53]. Recent approaches [66, 71] reduce 3D supervision by using LLMs [47, 48] to process reformatted text, but overlook critical visual cues – such as color, texture, viewpoint, and spatial layout – essential for precise localization (see Fig. 1).

To address these gaps, we present ***SeeGround***, a 3DVG framework that leverages 2D Vision-Language Models (VLMs) [16, 47, 57]. Trained on diverse image-text data, 2D-VLMs exhibit strong open-vocabulary understanding and transferability to zero-shot 3DVG [25, 74]. However, they cannot directly process 3D data. We thus propose a

cross-modal alignment representation that bridges 3D and 2D modalities by combining rendered images with spatially enriched text, enabling 2D-VLMs to reason over 3D scenes without task-specific retraining [35]. Specifically, we represent each 3D scene using a query-driven 2D rendering and structured 3D spatial descriptions. Compared to multi-view or bird's-eye strategies, our dynamic viewpoint captures both spatial context and object-level details. Meanwhile, the 3D text encodes accurate position and semantics from precomputed object detection, enhancing grounding performance. To associate image features with spatial descriptions, we introduce a **visual prompting** mechanism that highlights relevant objects in the rendered view. This helps the 2D-VLM resolve ambiguities in multi-object scenes by linking textual references to specific image regions, improving alignment and reducing distraction.

We evaluate our method on two benchmarks: it surpasses previous zero-shot baselines by 7.7% on *ScanRefer* [6] and 7.1% on *Nr3D* [1], while approaching fully supervised models. Moreover, it remains robust under incomplete textual input, leveraging visual cues for localization. To sum up, the contributions of this work are as follows:

- We propose ***SeeGround***, a training-free framework for zero-shot 3DVG, reformulating 3D scenes into 2D-VLM-compatible inputs using rendered views and spatial text.
- We design a query-guided viewpoint selection strategy to capture object-specific details and spatial context.
- We introduce a visual prompting technique to align 2D image features with 3D spatial descriptions, reducing grounding ambiguity in cluttered scenes.
- Our method achieves state-of-the-art zero-shot performance on *ScanRefer* and *Nr3D*, demonstrating strong generalization without 3D-specific training.

2. Related Work

3D Visual Grounding. Supervised 3DVG methods, such as ScanRefer [6] and ReferIt3D [1], align 3D scenes with language via attention-based frameworks like 3DVG-Transformer [73]. Recent works enhance multimodal fusion: ViewRefer [15] integrates LLMs for semantic-rich text; MVT [20] and LAR [2] use multi-view spatial cues; SAT [68] introduces 2D-assisted learning; and BUTD-DETR [23], ConcreteNet [56], and WS-3DVG [59] adopt transformer-based or weakly supervised frameworks. PQ3D [75] unifies multiple 3D-VL tasks. Despite strong benchmark performance, these methods require heavy annotations. Zero-shot approaches like LLM-Grounder [66] and ZSVG3D [71] avoid this but often miss visual cues vital for precise grounding.

3D Open-Vocabulary Understanding. Recent works enable open-vocabulary 3D understanding via 2D-3D alignment [8, 9, 39, 43, 65]. OpenScene [49] projects 2D CLIP features into 3D for segmentation; LeRF [26] fuses CLIP with NeRFs. OVIR-3D [44] and Agent3D-Zero [72] enhance instance retrieval and QA via multi-view reasoning. Other approaches, like RegionPLC [67], OpenMask3D [54], and OpenIns3D [22], rely on image-driven cues. SAI3D [69] uses 2D Semantic-SAM masks with 3D graph fusion. These works highlight the utility of 2D features in enriching 3D open-vocabulary tasks.

MLLMs for 3D Perception. Multimodal LLMs (MLLMs) extend 2D language grounding to 3D understanding [27, 38, 63, 64]. Scene-LLM [13] and Uni3DL [36] support captioning and segmentation in 3D scenes. 3D-ViSTA [74] and ConceptFusion [24] align 3D data with language through transformers and concept-level knowledge. GLOVER [46] enables 3D manipulation via language, while SceneVerse [25] provides annotated 3D environments for spatial learning. RLHF-V [52] further supports task planning via natural language. These advances show the potential of MLLMs in 3D tasks. Our work builds upon this by enabling zero-shot 3D grounding through multimodal alignment.

3. Methodology

Overview. The 3D visual grounding (3DVG) task aims to localize a target object in a 3D scene $\mathcal{S}$ based on a textual query Q, by predicting its 3D bounding box: $\mathbf{bbox} = \mathbf{3DVG}(\mathcal{S}, \mathrm{Q})$. We propose a novel 3DVG framework that integrates 2D vision-language models (2D-VLMs) with spatially enriched 3D representations. As conventional 3D formats are incompatible with 2D-VLM inputs, we introduce a **hybrid representation** that combines 2D rendered images and text-based 3D spatial descriptions. This enables 2D-VLMs to jointly reason over visual and spatial information without 3D-specific retraining.

Our approach comprises three modules: (1) a multimodal 3D representation (Sec. 3.1); (2) a Perspective Adaptation Module (Sec. 3.2); and (3) a Fusion Alignment Module (Sec. 3.3). This design allows the 2D-VLM to accurately interpret and localize objects in complex 3D scenes. Fig. 2 provides an overview of our framework.

3.1. Multimodal 3D Representation

We leverage 2D-VLMs pretrained on large-scale image-text data to enable open-set understanding of novel objects and scenes. However, conventional 3D representations – point clouds [17, 49], voxels [37], and implicit fields [26] – are incompatible with 2D-VLM input. To bridge this gap, we propose a hybrid representation that combines 2D rendered images with text-based 3D spatial descriptions.

Text-based 3D Spatial Descriptions. We first detect all objects in the scene using an open-vocabulary 3D detector: $(\mathbf{bbox}, \mathbf{sem})_{i=1}^{N} = \mathbf{OVDet}(\mathcal{S})$, where $\mathbf{bbox}$ and $\mathbf{sem}$ are each object's 3D bounding box and semantic label. These are converted to text and stored in an object lookup table (OLT) for reuse: $\mathcal{OLT} = \{(\mathbf{bbox}, \mathbf{sem})\}_{i=1}^{N}$. Here,

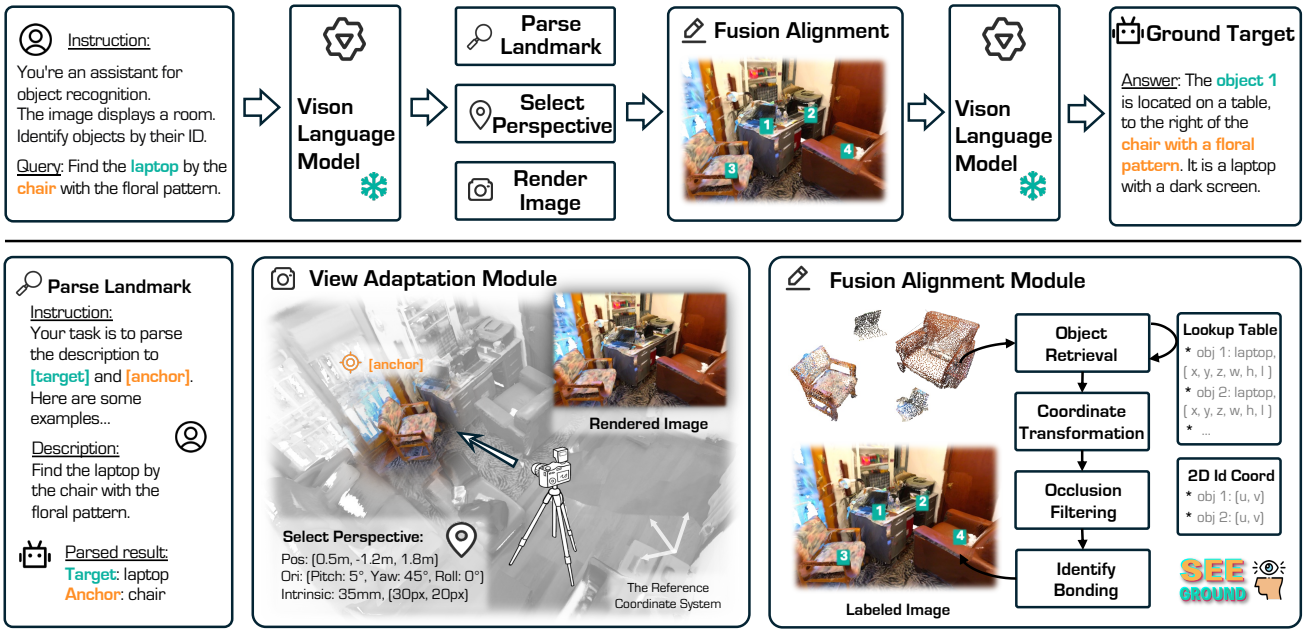


Figure 2. Overview of the *SeeGround* framework. We first use a 2D-VLM to interpret the query, identifying both the target object (*e.g.*, "laptop") and a context-providing anchor (*e.g.*, *"chair with a floral pattern"*). A dynamic viewpoint is then selected based on the anchor's position, enabling the capture of a 2D rendered image that aligns with the query's spatial requirements. Using the Object Lookup Table ($\mathcal{OLT}$), we retrieve the 3D bounding boxes of relevant objects, project them onto the 2D image, and apply visual prompts to mark visible objects, filtering out occlusions. The image with prompts, along with the spatial descriptions and query, is then input into the 2D-VLM for precise localization of the target object. Finally, the 2D-VLM outputs the target object's ID, which is then used to retrieve its 3D bounding box from the $\mathcal{OLT}$, providing the final, accurate 3D position in the scene.

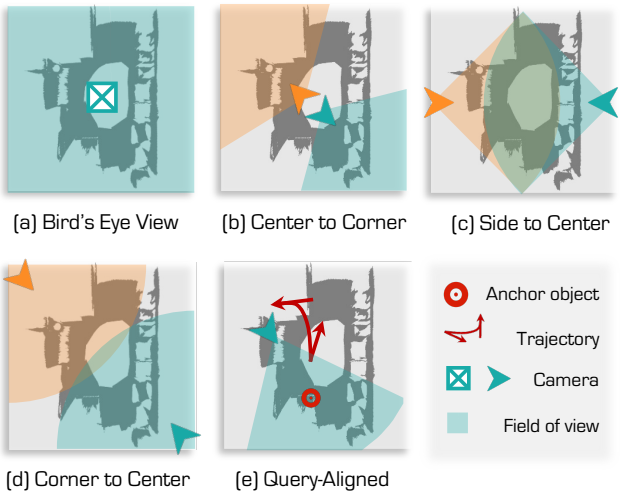


Figure 3. Illustrative example of different perspective selection strategies. Our "Query-Aligned" method dynamically adapts the viewpoint to match the spatial context of the query, enhancing detail and relevance of visible objects compared to static methods.

$\mathcal{OLT}$ supports efficient spatial reasoning and avoids redundant computation across multiple queries.

Hybrid 3D Scene Representation. Text captures 3D layout and semantics, but lacks fine visual cues. To complement this, we render a 2D image aligned with the query: $(\mathbf{I}, \mathcal{T}) = \mathbf{F}(\mathcal{S}, \mathbf{Q}, \mathcal{OLT})$, where $\mathbf{I}$ is the rendered image and $\mathcal{T}$ is the textual spatial description. Combined, $\mathbf{I}$ and $\mathcal{T}$ allow 2D-VLM to access both visual appearance (*e.g.*, color, texture, shape) and accurate 3D spatial semantics, enabling comprehensive scene understanding.

3.2. Perspective Adaptation Module

Existing view selection strategies often misalign with the query's perspective. For example, LAR [2] renders object-centric multi-views but lacks scene context, while bird's-eye view (BEV) provides global coverage but loses height information, leading to occlusions (see Fig. 3(a)). Multi-view or multi-scale methods [22] expand coverage (see Fig. 3(b)-(d)), but still rely on fixed viewpoints. Moreover, 2D-VLMs often misinterpret scenes when the rendered view does not reflect the query perspective. To address this, we propose a query-driven dynamic rendering method that aligns the viewpoint with the query, capturing more relevant details (see Fig. 3(e)).

Dynamic Perspective Selection. Given a query Q, the 2D-VLM extracts an anchor object $\boldsymbol{A}$ and candidate targets $\mathcal{O}^{(C)}$ using example prompts $\mathcal{E}^{(E)}$: $\left(\boldsymbol{A}, \mathcal{O}^{(C)}\right) = \text{VLM}\left(\mathsf{Q}, \mathcal{E}^{(E)}\right)$. The viewpoint is placed at the scene center, facing $\boldsymbol{A}$, and then shifted backward and upward to ex-

Table 1. Evaluations of 3DVG on *ScanRefer* [6] validation set. Results are reported for *"Unique"* (scenes with a single target object) and *"Multiple"* (scenes with distractors of the same class) subsets, along with overall performance. * indicates results on selected 250 samples.

Method	Venue	Supervision	Agent	Unique		Multiple		Overall	
				Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [6]	ECCV'20	Fully	-	67.6	46.2	32.1	21.3	39.0	26.1
InstanceRefer [70]	ICCV'21	Fully	-	77.5	66.8	31.3	24.8	40.2	32.9
3DVG-T [73]	ICCV'21	Fully	-	77.2	58.5	38.4	28.7	45.9	34.5
BUTD-DETR [23]	ECCV'22	Fully	-	84.2	66.3	46.6	35.1	52.2	39.8
EDA [61]	CVPR'23	Fully	-	85.8	68.6	49.1	37.6	54.6	42.3
3D-VisTA [74]	ICCV'23	Fully	-	81.6	75.1	43.7	39.1	50.6	45.8
G3-LQ [58]	CVPR'24	Fully	-	88.6	73.3	50.2	39.7	56.0	44.7
MCLN [50]	ECCV'24	Fully	-	86.9	72.7	52.0	40.8	57.2	45.7
ConcreteNet [56]	ECCV'24	Fully	-	86.4	82.1	42.4	38.4	50.6	46.5
WS-3DVG [59]	ICCV'23	Weakly	-	-	-	-	-	27.4	22.0
LERF [26]	ICCV'23	Zero-Shot	CLIP [51]	-	-	-	-	4.8	0.9
OpenScene [49]	CVPR'23	Zero-Shot	CLIP [51]	20.1	13.1	11.1	4.4	13.2	6.5
LLM-G [66]	ICRA'24	Zero-Shot	GPT-3.5 [48]	-	-	-	-	14.3	4.7
LLM-G [66]	ICRA'24	Zero-Shot	GPT-4 turbo [47]	-	-	-	-	17.1	5.3
ZSVG3D [71]	CVPR'24	Zero-Shot	GPT-4 turbo [47]	63.8	58.4	27.7	24.6	36.4	32.7
VLM-Grounder* [62]	CoRL'24	Zero-Shot	GPT-4V [47]	66.0	29.8	**48.3**	**33.5**	**51.6**	32.8
SeeGround	**Ours**	Zero-Shot	Qwen2-VL-72b [57]	**75.7**	**68.9**	34.0	30.0	44.1	**39.4**

Table 2. Performance on *Nr3D* [1] validation set. Queries are labeled as *"Easy"* (one distractor) or *"Hard"* (multiple distractors), and as *"View-Dependent"* or *"View-Independent"* based on viewpoint requirements for grounding.

Method	Easy	Hard	Dep.	Indep.	Overall
Supervision: Fully Supervised					
ReferIt3DNet [1]	43.6	27.9	32.5	37.1	35.6
TGNN [19]	44.2	30.6	35.8	38.0	37.3
InstanceRefer [70]	46.0	31.8	34.5	41.9	38.8
3DVG-T [73]	48.5	34.8	34.8	43.7	40.8
BUTD-DETR [23]	60.7	48.4	46.0	58.0	54.6
MiKASA [5]	69.7	59.4	65.4	64.0	64.4
ViL3DRel [10]	70.2	57.4	62.0	64.5	64.4
Supervision: Weakly Supervised					
WS-3DVG [59]	27.3	18.0	21.6	22.9	22.5
Supervision: Zero-Shot					
ZSVG3D [71]	46.5	31.7	36.8	40.0	39.0
SeeGround	**54.5**	**38.3**	**42.3**	**48.2**	**46.1**

pand coverage. If no anchor is found (*e.g.*, query describes multiple similar objects), a placeholder anchor is used – set to the centroid of $\mathcal{O}^{(C)}$ – and the same strategy applies.

Query-Aligned Image Rendering. Given the selected viewpoint, the function look-at-view-transform computes rotation $\mathbf{R}_c$ and translation $\mathbf{T}_c$ relative to $\boldsymbol{A}$. The 2D rendered image is obtained via: $\mathbf{I} = \text{Render}(\mathcal{S}, \mathbf{R}_c, \mathbf{T}_c)$. This generates a query-aligned image that preserves relevant visual details while filtering distractions, thereby improving 2D-VLM localization accuracy (see Fig. 3(e)).

3.3. Fusion Alignment Module

Although 2D images and spatial descriptions offer rich cues, directly feeding them into 2D-VLM may fail to associate visual features with 3D semantics, especially in scenes with similar objects (*e.g.*, multiple chairs), leading to grounding errors. We propose a Fusion Alignment Module to explicitly align 2D visual features with text descriptions.

Depth-Aware Visual Prompting. Given the rendered image $\mathbf{I}$, we retrieve the 3D points of each object in the scene from $\mathcal{OLT}$, then project them to 2D using $\mathbf{R}_c$ and $\mathbf{T}_c$. To address occlusion, we compare point depths with the depth map and retain only visible points. For each object o, a visual prompt $\mathcal{M}_o$ is placed at the center of its visible projections. The prompted image is computed as: $\mathbf{I}_m = \mathbf{I} \odot \left(1 - \mathbb{1}_{\mathcal{P}_{\text{visible}}(o)}\right) + \mathcal{M}_o \odot \mathbb{1}_{\mathcal{P}_{\text{visible}}(o)}$, where $\mathbb{1}_{\mathcal{P}_{\text{visible}}(o)}$ indicates visible pixels of object o.

Object Prediction with 2D-VLM. Finally, given the query Q, prompted image $\mathbf{I}_m$, and spatial text $\mathcal{T}$, the 2D-VLM predicts the referred object: $\hat{o} = \text{VLM}\left(\mathrm{Q} \,|\, \mathbf{I}_m, \mathcal{T}\right)$. By explicitly aligning visual and spatial cues, this module reduces ambiguity and enhances localization in cluttered scenes.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate on two standard 3DVG benchmarks. **ScanRefer** [6] offers $51,500$ descriptions across 800 ScanNet scenes, with queries labeled as *"Unique"* or *"Multiple"* based on distractor presence. **Nr3D** [1], part of ReferIt3D, contains $41,503$ two-player game queries, split into *"Easy"*/*"Hard"* (by distractor count) and *"View-Dependent"*/*"View-Independent"* (by viewpoint reliance). ScanRefer emphasizes sparse point cloud grounding, while Nr3D provides full 3D box annotations.

Implementation Details. We use Qwen2-VL-72B [57] as the backbone VLM. Ablation studies are conducted on the Nr3D validation set. Images are rendered at 1000×1000 resolution, excluding the top $0.3\,\text{m}$ to align with closed-

room setups. For object detection, we follow ZSVG3D [71] and use Mask3D [54] to ensure consistent evaluation.

4.2. Comparative Study

On **ScanRefer**, we achieve 75.7% / 68.9% (Acc@0.25/0.5) on *"Unique"*, and 34.0% / 30.0% on *"Multiple"*, surpassing all zero-shot and weakly supervised baselines [59, 66, 71], and approaching fully supervised methods [50, 56]. On **Nr3D**, it attains 46.1% overall accuracy, outperforming the previous zero-shot SOTA by +7.1% [71], and remains robust across *"Easy"*/*"Hard"* (54.5% / 38.3%) and *"View-Dependent"*/*"View-Independent"* (42.3% / 48.2%) subsets, closing the gap with fully supervised methods [23].

4.3. Ablation Study

Effect of Architecture Design. We begin by evaluating the effectiveness of each module in the proposed architecture. The experimental results are presented in Tab. 3.

- *Layout of the Scene.* Using only 3D coordinates (37.7%, Tab. 3(a)) provides the basic location of objects but achieves low accuracy. Adding layout (39.7%, Tab. 3(b)), which renders 3D boxes in 2D without color/texture, improves accuracy by providing spatial context that helps the model understand object positions and sizes.
- *Visual Clues.* We find that adding color/texture (39.5%, Tab. 3 (c)) helps the model distinguish between similar objects, like "the white keyboard" versus "the black keyboard" (Fig. 4 (a)). This setup tends to improve accuracy over layout alone by offering object-specific visual cues.
- *Fusion Alignment Module.* As shown in Tab. 3 (d), our proposed Fusion Alignment Module provides a significant increase in accuracy (43.3%) by associating 2D images with text descriptions.
- *Perspective Adaptation Module.* Perspective Adaptation Module (45.0%, Tab. 3 (e)) further improves accuracy by aligning the scene's viewpoint with the query's spatial context (Fig. 4 (b)), helping the model understand the positional context and reducing ambiguity.
- *Full Configuration.* We observe that the highest accuracy (46.1%) is achieved with the full configuration (Tab. 3 (f)). This further validates the effectiveness and efficiency of the proposed SeeGround framework.

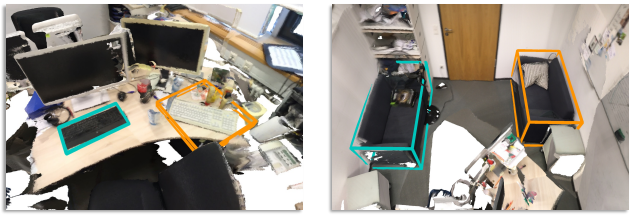
Ours *vs*. Prior Art. ZSVG3D [71] projects object centers onto a 2D image and uses predefined functions to infer spatial relations, but this approach lacks flexibility, omits visual cues, and ignores contextual objects, risking misidentification if detection fails (Fig. 6). Fig. 5a compares the VLM version of ZSVG3D's projection, showing only target and anchor centers, with no background or visual detail. In contrast, our method captures full images, and allows inference of undetected objects via contextual cues, as in Fig. 6.

Qwen2-VL *vs*. GPT-4. To ensure wider applicability, cost-effectiveness, and reproducibility, we use the open-source model Qwen2-VL [57] in our method. To ensure fairness, we re-evaluate ZSVG3D [71] with Qwen2-VL instead of GPT-4 [47], as shown in Fig. 5b, enabling direct comparison with our method. Using the same model, our approach outperforms ZSVG3D across all difficulty levels, confirming its effectiveness independently of model choice. We use ZSVG3D's program generation prompt with Qwen2-VL, keeping other steps identical.

Effect of View Selection Strategy. Tab. 4 highlights the benefits of our query-driven perspective alignment (see Fig. 3) over static strategies. Fixed views – Center2Corner, Edge2Center, Corner2Center – lack adaptability, while Bird's Eye View, though global, misses critical spatial cues

Table 3. **Ablation study** on different components in our framework on *Nr3D* [1]. *"3D Pos."*: 3D object coordinates; *"Layout"*: Scene layout; *"Texture"*: Object color/texture; *"FAM"*: Fusion Alignment Module; and *"PAM"*: Perspective Adaptation Module.

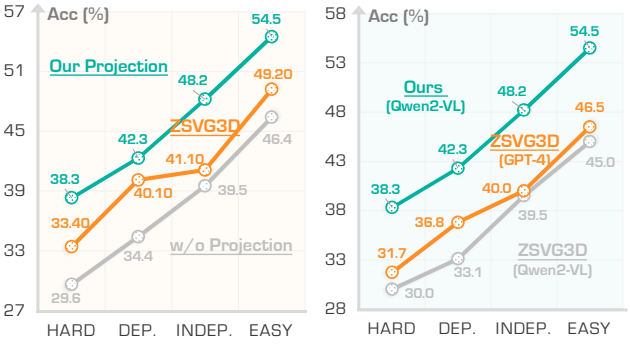
#	3D Pos.	Layout	Texture	FAM	PAM	Overall
(a)	✓	✗	✗	✗	✗	37.7
(b)	✓	✓	✗	✗	✗	39.7
(c)	✓	✗	✓	✗	✗	39.5
(d)	✓	✓	✓	✓	✗	43.3
(e)	✗	✓	✓	✓	✓	45.0
(f)	✓	✓	✓	✓	✓	**46.1**



(a) the black keyboard, not white, that is place on the table.

(b) When you are facing the door, it's the couch on the left.

Figure 4. **Qualitative Results.** Rendered images are presented, including the incorrectly identified objects (**Orange**) and correctly identified objects (**Green**). Key visual cues are underlined.



(a) Projection Method (b) Language Agent

Figure 5. **Ablation study** on using (a) different projection methods (ours *vs*. ZSVG3D [71]); and (b) different language agents (GPT-4 [47] *vs*. Qwen2-VL [57]). The results are from *Nr3D* [1].

Table 4. Performance comparison of different perspective selection strategies. Our method results in consistently higher accuracy across all difficulty levels on Nr3D [1] validation set.

Type	Easy	Hard	Dep.	Indep.	Overall
Center2Corner	49.5	31.4	35.1	42.9	40.2
Edege2Center	51.0	32.7	36.6	44.2	41.5
Corner2Center	49.8	33.4	35.5	44.5	41.3
Bird's Eye View	53.4	33.9	36.9	46.8	43.3
Query-aligned	**54.5**	**38.3**	**42.3**	**48.2**	**46.1**

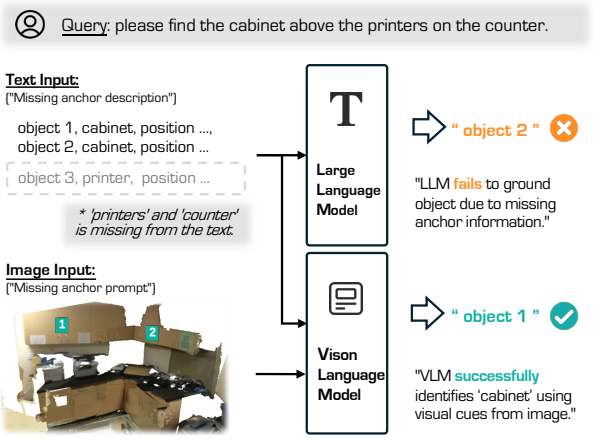
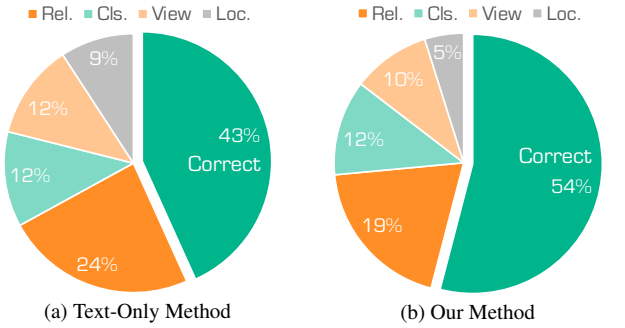


Figure 6. An example of the robustness of the proposed framework in identifying the 'cabinet' by leveraging visual context, even when key information ('printers' and 'counter') is missing from text input – an issue that commonly arises in scenarios with detection errors or omissions. Our method is more robust than prior art.

like orientation and height. Our dynamic strategy yields consistent gains, notably on *"Hard"* (+4.4%) and *"Dependent"* (+5.7%) queries, demonstrating the importance of flexible, context-aware viewpoint selection for 3DVG.

Robustness Evaluation with Incomplete Textual Descriptions. As shown in Fig. 6, we tested our approach's robustness with incomplete textual information, simulating common misdetection scenarios. By omitting an anchor object from the text while retaining the target, our model uses visual cues to compensate, achieving accurate localization. In contrast, LLM performance degrades without the anchor. These results demonstrate that our method maintains high accuracy with partial text, underscoring the importance of integrating visual and textual data for more reliable 3DVG.

Type-Wise Error Analysis. We analyzed 185 randomly sampled cases from 10 rooms to identify common failure modes (Fig. 7). The reduction in localization and identification errors underscores the benefit of visual input for spatial understanding. However, errors involving spatial relationships remain frequent (19%), suggesting challenges in precise spatial reasoning. Incorporating dedicated reasoning modules may help. Our current viewpoint selection also struggles with complex references like *"when the window is*



(a) Text-Only Method (b) Our Method

Figure 7. Error distributions between the Text-Only Method (a) and Our Method (b), based on four error types: Relation Errors (**Rel.**, spatial relationship misunderstandings like "next to" or "on the corner"), Classification Errors (**Cls.**, object category misidentifications), Viewpoint Errors (**View**, errors in interpreting specific observation viewpoints), and Localization Errors (**Loc.**, errors in pinpointing the target object within the scene).

on the left" or *"upon entering from the door"*. Additionally, rendering quality affects object distinction; due to reliance on dataset-provided point clouds, our renderings lack fine texture and boundary details. Future improvements may include higher-fidelity rendering for clearer visual cues.

5. Conclusion

In this paper, we presented ***SeeGround***, a zero-shot 3D visual grounding framework that bridges 3D data and 2D VLMs via query-aligned rendered images and spatial descriptions. Our Perspective Adaptation Module aligns views with the query, while the Fusion Alignment Module integrates visual and spatial cues for robust localization. Experiments on the ScanRefer and Nr3D datasets show that our method outperforms prior zero-shot methods and rivals supervised approaches without 3D-specific training.

References

[1] Panos Achlioptas et al. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, pages 422–440, 2020.

[2] Eslam Mohamed Bakr et al. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. In *NeurIPS*, pages 37146–37158, 2022.

[3] Jens Behley et al. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, pages 9297–9307, 2019.

[4] Hengwei Bian et al. Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. *arXiv preprint arXiv:2410.18084*, 2024.

[5] Chun-Peng Chang et al. Mikasa: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *CVPR*, pages 14131–14140, 2024.

[6] Dave Zhenyu Chen et al. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221, 2020.

[7] Runnan Chen et al. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *CVPR*, pages 7020–7030, 2023.

[8] Runnan Chen et al. Towards label-free scene understanding by vision foundation models. In *NeurIPS*, pages 75896–75910, 2023.

[9] Runnan Chen et al. Ovgaussian: Generalizable 3d gaussian segmentation with open vocabularies. *arXiv preprint arXiv:2501.00326*, 2025.

[10] Shizhe Chen et al. Language conditioned spatial relation reasoning for 3d object grounding. In *NeurIPS*, 2022.

[11] Shizhe Chen et al. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, pages 16537–16547, 2022.

[12] Whye Kit Fong et al. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *RA-L*, 7:3795–3802, 2022.

[13] Rao Fu et al. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

[14] Zeying Gong et al. From cognition to precognition: A future-aware framework for social navigation. *arXiv preprint arXiv:2409.13244*, 2024.

[15] Zoey Guo et al. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *ICCV*, pages 15372–15383, 2023.

[16] Wenyi Hong et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[17] Yining Hong et al. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, pages 20482–20494, 2023.

[18] Tianshuai Hu et al. Dhp-mapping: A dense panoptic mapping system with hierarchical world representation and label optimization techniques. In *IROS*, pages 1101–1107, 2024.

[19] Pin-Hao Huang et al. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*, pages 1610–1618, 2021.

[20] Shijia Huang et al. Multi-view transformer for 3d visual grounding. In *CVPR*, pages 15524–15533, 2022.

[21] Zanming Huang et al. Assister: Assistive navigation via conditional instruction generation. In *ECCV*, pages 271–289, 2022.

[22] Zhening Huang et al. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *ECCV*, pages 169–185, 2025.

[23] Ayush Jain et al. Bottom up top down detection transformers for language grounding in images and point clouds. In *ECCV*, pages 417–433, 2022.

[24] Krishna Murthy Jatavallabhula et al. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems*, 2023.

[25] Baoxiong Jia et al. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *ECCV*, pages 289–310, 2025.

[26] Justin Kerr et al. Lerf: Language embedded radiance fields. In *ICCV*, pages 19729–19739, 2023.

[27] Lingdong Kong et al. Lasermix for semi-supervised lidar semantic segmentation. In *CVPR*, pages 21705–21715, 2023.

[28] Lingdong Kong et al. Rethinking range view representation for lidar segmentation. In *ICCV*, pages 228–240, 2023.

[29] Lingdong Kong et al. Robo3d: Towards robust and reliable 3d perception against corruptions. In *ICCV*, pages 19994–20006, 2023.

[30] Lingdong Kong et al. Calib3d: Calibrating model preferences for reliable 3d scene understanding. In *WACV*, pages 1965–1978, 2025.

[31] Lingdong Kong et al. Multi-modal data-efficient 3d scene understanding for autonomous drivin. *TPAMI*, 47(5):3748–3765, 2025.

[32] Lei Lai et al. Xvo: Generalized visual odometry via cross-modal self-training. In *ICCV*, pages 10094–10105, 2023.

[33] Rong Li et al. Coarse3d: Class-prototypes for contrastive learning in weakly-supervised 3d point cloud segmentation. *arXiv preprint arXiv:2210.01784*, 2022.

[34] Rong Li et al. Tfnet: Exploiting temporal cues for fast and accurate lidar semantic segmentation. In *CVPR*, pages 4547–4556, 2024.

[35] Rong Li et al. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. *arXiv preprint arXiv:2412.04383*, 2024.

[36] Xiang Li et al. Uni3dl: A unified model for 3d and language understanding. *arXiv preprint arXiv:2312.03026*, 2023.

[37] Ye Li et al. Is your lidar placement optimized for 3d scene understanding? In *NeurIPS*, pages 34980–35017, 2024.

[38] Youquan Liu et al. Segment any point cloud sequences by distilling vision foundation models. In *NeurIPS*, pages 37193–37229, 2023.

[39] Youquan Liu et al. Multi-space alignments towards universal lidar segmentation. In *CVPR*, pages 14648–14661, 2024.

[40] Zhuoman Liu et al. Deep view synthesis via self-consistent generative network. *IEEE Transactions on Multimedia*, 24: 451–465, 2021.

[41] Zhuoman Liu et al. Raydf: neural ray-surface distance fields with multi-view consistency. *arXiv preprint arXiv:2310.19629*, 2023.

[42] Zhuoman Liu et al. Unleashing the potential of multi-modal foundation models and video diffusion for 4d dynamic physical scene simulation. *arXiv preprint arXiv:2411.14423*, 2024.

[43] Dong Lu et al. Geal: Generalizable 3d affordance learning with cross-modal consistency. *arXiv preprint arXiv:2412.09511*, 2025.

[44] Shiyang Lu et al. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *CoRL*, pages 1610–1620, 2023.

[45] Teli Ma et al. An examination of the compositionality of large generative vision-language models. *arXiv preprint arXiv:2308.10509*, 2023.

[46] Teli Ma et al. Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping. *arXiv preprint arXiv:2411.12286*, 2024.

[47] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[48] Long Ouyang et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022.

[49] Songyou Peng et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023.

[50] Zhipeng Qian et al. Multi-branch collaborative learning network for 3d visual grounding. In *ECCV*, pages 381–398, 2025.

[51] Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[52] Lingfeng Sun et al. Interactive planning using large language models for partially observable robotic tasks. In *ICRA*, pages 14054–14061, 2024.

[53] Pei Sun et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020.

[54] Ayça Takmaz et al. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

[55] Mingkui Tan et al. Epmf: Efficient perception-aware multi-sensor fusion for 3d semantic segmentation. *TPAMI*, 46(12): 8258–8273, 2024.

[56] Ozan Unal et al. Four ways to improve verbo-visual fusion for dense 3d visual grounding. In *ECCV*, pages 196–213, 2025.

[57] Peng Wang et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[58] Yuan Wang et al. G3-lq: Marrying hyperbolic alignment with explicit semantic-geometric modeling for 3d visual grounding. In *CVPR*, pages 13917–13926, 2024.

[59] Zehan Wang et al. Distilling coarse-to-fine semantic matching knowledge for weakly supervised 3d visual grounding. In *ICCV*, pages 2662–2671, 2023.

[60] Xiaokang Wei et al. Sir: Multi-view inverse rendering with decomposable shadow for indoor scenes. *arXiv preprint arXiv:2402.06136*, 2024.

[61] Yanmin Wu et al. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *CVPR*, pages 19231–19242, 2023.

[62] Runsen Xu et al. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024.

[63] Xiang Xu et al. 4d contrastive superflows are dense 3d representation learners. In *ECCV*, pages 58–80, 2024.

[64] Xiang Xu et al. Frnet: Frustum-range networks for scalable lidar segmentation. *TIP*, 34:2173–2186, 2025.

[65] Xiang Xu et al. Limoe: Mixture of lidar representation learners from automotive scenes. *arXiv preprint arXiv:2501.04004*, 2025.

[66] Jianing Yang et al. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *ICRA*, pages 7694–7701, 2024.

[67] Jihan Yang et al. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *CVPR*, pages 19823–19832, 2024.

[68] Zhengyuan Yang et al. Sat: 2d semantics assisted training for 3d visual grounding. In *ICCV*, pages 1856–1866, 2021.

[69] Yingda Yin et al. Sai3d: Segment any instance in 3d scenes. In *CVPR*, pages 3292–3302, 2024.

[70] Zhihao Yuan et al. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *ICCV*, pages 1791–1800, 2021.

[71] Zhihao Yuan et al. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *CVPR*, pages 20623–20633, 2024.

[72] Sha Zhang et al. Agent3d-zero: An agent for zero-shot 3d understanding. In *arXiv preprint arXiv:2403.11835*, 2024.

[73] Lichen Zhao et al. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, pages 2928–2937, 2021.

[74] Ziyu Zhu et al. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, pages 2911–2921, 2023.

[75] Ziyu Zhu et al. Unifying 3d vision-language understanding via promptable queries. In *ECCV*, pages 188–206, 2024.

[76] Zhuangwei Zhuang et al. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *ICCV*, pages 16280–16290, 2021.

[77] Zhuangwei Zhuang et al. Robust 3d semantic occupancy prediction with calibration-free spatial transformation. *arXiv preprint arXiv:2411.12177*, 2024.