VERA: VISUAL EVIDENCE REFINEMENT AGENT FOR TRAINING-FREE 3D REFERRING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in multimodal large language models (MLLMs) have substantially improved visual understanding and reasoning in 2D domains, opening new possibilities for 3D scene understanding without task-specific training. However, extending these capabilities to 3D referring segmentation (3D-RES) remains challenging, as it requires not only fine-grained understanding of appearance and spatial relationships, but also accurate alignment between language semantics and point-level information in 3D scenes. Consequently, achieving reliable 3D-RES without task-specific supervision remains difficult. To address these challenges, we propose VERA, a training-free framework for 3D-RES that integrates a pretrained MLLM with complementary 2D and 3D perception tools. Specifically, a 3D Candidate Generation module first constructs a query-relevant instance candidate space while preserving the corresponding 3D mask for each candidate. Building upon this candidate space, we introduce a Visual Evidence Self-Refinement module that formulates target selection as a visual evidence-driven process of hypothesis verification and self-refinement. Through this process, the module jointly leverages global scene views, local contextual observations, and object-level crops to assess the semantic and spatial consistency between candidate hypotheses and the referring expression. When the initial hypothesis lacks sufficient visual support, verification feedback is used to reassess competing candidates and refine the current hypothesis through direct visual comparison. The point-level mask associated with the selected candidate is returned as the final segmentation. Extensive experiments on multiple benchmarks demonstrate the effectiveness of VERA for training-free 3D-RES without task-specific supervision.

1 INTRODUCTION

3D referring segmentation (3D-RES) aims to locate the target instance specified by a natural language expression in a 3D scene and predict its point-level mask, with broad applications in human-computer interaction (De La Torre et al., 2024; Chen et al., 2025a) and virtual reality (Chen et al., 2025b; Wang et al., 2025). Existing 3D-RES approaches typically follow two paradigms: task-specific training (Chen et al., 2025c; Wang et al., 2026b; Qian et al., 2024) and supervised fine-tuning of multimodal large language models (MLLMs) (Fan et al., 2026; Chen et al., 2024; Wu et al., 2026b). The former learns from paired 3D scenes, referring expressions, and instance-level mask annotations, whereas the latter adapts pretrained MLLMs using annotated 3D data. Although effective, both paradigms require task-specific supervision, limiting their scalability to diverse scenes and referring expressions.

Recent advances in multimodal agents have demonstrated the ability of large models to reason through interaction with external tools and iterative information (Wu et al., 2023; Gao et al., 2025; Yang et al., 2025; Jiang et al., 2026). Building upon these capabilities, recent studies have explored their applications in 3D scene understanding, including 3D visual question answering, language-guided scene reasoning, and object-level grounding (Yang et al., 2024; Azuma et al., 2022; Chen et al., 2026; Jin et al., 2026). These studies demonstrate that pretrained models can interact with external perception tools to actively acquire task-relevant information and perform multi-step reasoning in complex 3D environments, as shown in Fig. 1 (a) and (b). However, existing agentic paradigms mainly focus on scene understanding or object localization, while extending them to 3D referring segmentation with instance identification and point-level masks remains unexplored.

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.

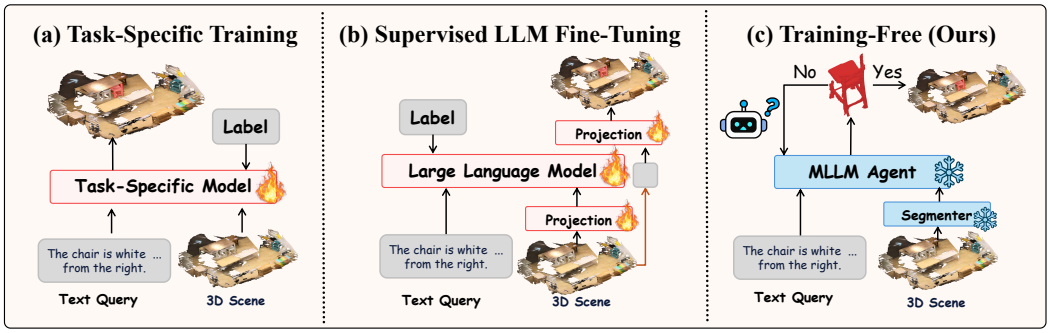

Figure 1: Comparison of different paradigms for 3D-RES. (a) Task-specific training with 3D annotations. (b) Supervised fine-tuning of pretrained models with labeled data. (c) Our training-free paradigm using pretrained multimodal models and off-the-shelf tools.

Realizing training-free 3D-RES with an agent remains challenging. First, efficiently acquiring task-relevant evidence from complex 3D scenes is difficult. Exhaustively examining all instances incurs substantial computational costs, whereas overly aggressive filtering based on incomplete semantic cues may discard the target due to perception errors and ambiguous descriptions. Second, reliably verifying candidate hypotheses is challenging because multiple instances may satisfy different subsets of the referring constraints. The agent must determine whether the current hypothesis is adequately supported or whether further evidence and refinement are necessary.

To address these challenges, we propose VERA, a visual evidence refinement agent for training-free 3D-RES, as shown in Fig. 1 (c). VERA integrates a pretrained MLLMs with complementary 2D and 3D perception tools and formulates referring segmentation as an evidence-driven hypothesis verification and refinement process. Specifically, VERA first constructs a query-relevant candidate space with preserved point-level masks through the 3D Candidate Generation module. It then performs multi-granularity visual evidence verification through the Visual Evidence Self-Refinement module to evaluate candidate hypotheses and refine predictions when stronger evidence is identified.

Our contributions are summarized as follows:

- We introduce VERA, a training-free agentic framework for 3D-RES that leverages pretrained MLLMs with external perception tools to perform evidence-guided target selection without task-specific supervision.
- We propose a visual evidence-based refinement paradigm that formulates 3D-RES as hypothesis verification, where multi-granularity observations are used to verify and refine candidate instances.
- Extensive experimental evaluations are conducted on multiple 3D-RES benchmarks. The results demonstrate the effectiveness of VERA and validate the contributions of candidate generation and visual evidence.

2 Related Work

Fully supervised 3D referring expression segmentation. 3D-RES predicts point-level masks of the referred instance from language queries. Existing supervised approaches (Wu et al., 2026a; Huang et al., 2021; Qian et al., 2024; Wu et al., 2024c; He & Ding, 2024; Li et al., 2026) address this task through cross-modal alignment, proposal matching, and geometric reasoning. Recent extensions further explore enhanced appearance modeling, spatial relation reasoning, and multi-target referring segmentation (Chen et al., 2025c; Wang et al., 2026b; Wu et al., 2024a; Huang & Wang, 2026; Wu et al., 2024b). Although these methods achieve strong performance, their capabilities are obtained through task-specific supervision on annotated 3D referring segmentation datasets, requiring expensive data collection and model adaptation.

LLM-based 3D understanding and segmentation. Recent studies explore extending LLMs and MLLMs to 3D domains through feature alignment and instruction tuning, enabling general 3D scene

understanding and reasoning capabilities (Fan et al., 2026; Chen et al., 2024; Wu et al., 2026b; Xu et al., 2024b; Yang et al., 2026; Huang et al., 2025b; Liu et al., 2026). Beyond general scene understanding, several works investigate language-driven 3D segmentation and referring tasks. SegPoint (He et al., 2024), Reason3D (Huang et al., 2025a), and 3D-LLaVA (Deng et al., 2025) introduce large-model-based frameworks for point-cloud reasoning and mask prediction, while MLLM-For3D (Huang et al., 2026a) transfers 2D multimodal reasoning capabilities to 3D scenes through multi-view supervision. However, these approaches generally require learned 3D-language alignment, instruction tuning, or additional supervision, and therefore do not provide a training-free solution for point-cloud-based 3D referring segmentation without task-specific adaptation.

Zero-shot 3D understanding with MLLMs-based agents. Recent studies explore leveraging pretrained MLLMs for zero-shot 3D understanding by presenting 3D scenes through rendered visual observations (Zhang et al., 2024; Qi et al., 2026; Sun et al., 2025; Li et al., 2025). Agent3D-Zero (Zhang et al., 2024) demonstrates that MLLMs can reason about 3D environments from multi-view images. Following this direction, SeeGround (Li et al., 2025) and VLM-Grounder (Xu et al., 2024a) explore query-guided visual reasoning and multi-view projection for 3D understanding. Recent works further investigate evidence-driven reasoning mechanisms for 3D scenes. SPAZER (Jin et al., 2026) verifies candidate hypotheses using 3D views and local visual evidence, while other approaches explore perception-planning feedback (Wang et al., 2026a) and tool-assisted 3D reasoning (Drozdov et al., 2026). Training-free 3D segmentation has been explored through scene representations, such as sparse voxels or 3D Gaussian representations (Huang et al., 2026b; Ding et al., 2026). However, a training-free framework that leverages pretrained MLLMs for point-cloud 3D referring segmentation with evidence-driven hypothesis verification remains underexplored, especially for maintaining correspondence between candidate instances and point-level masks.

3 METHOD

3.1 OVERVIEW

We propose VERA, a training-free framework that integrates a pretrained MLLM with complementary 2D and 3D perception tools. As illustrated in Fig. 2, VERA performs 3D-RES in two stages. The 3D Candidate Generation module constructs a query-relevant candidate space while preserving point-level masks. The Visual Evidence Self-Refinement module then verifies candidates with multi-granularity visual evidence and refines the prediction when necessary. The mask of the final verified candidate is returned as the segmentation output.

3.2 3D CANDIDATE GENERATION

To address the challenge of efficiently obtaining query-relevant candidates from complex 3D scenes, we introduce the 3D Candidate Generation module. Although MLLMs exhibit strong cross-modal reasoning capabilities, they remain limited in directly processing raw 3D point clouds and producing precise point-level masks. VERA therefore does not directly apply the MLLMs to the entire 3D scene. Instead, it first constructs a compact, query-relevant hypothesis space, where each candidate instance remains associated with its corresponding point-level mask for subsequent Visual Evidence Self-Refinement.

Given a 3D scene $\mathcal{S}$ and a referring expression Q, we represent the input point cloud as $\mathcal{S} = \{p_i\}_{i=1}^{N}$, where $p_i \in \mathbb{R}^6$ contains the 3D coordinates and RGB attributes of each point. A frozen off-the-shelf 3D perception backbone decomposes the scene into object-level instance proposals:

$$\mathcal{O} = \{O_i\}_{i=1}^{M} = \Phi_{3D}(\mathcal{S}), \tag{1}$$

where $\Phi_{3D}(\cdot)$ denotes the pretrained 3D instance segmentation function and $\mathcal{O}$ denotes the proposal pool. Each proposal O_i is associated with a point-level mask M_i, and the pair (O_i, M_i) is preserved throughout candidate generation and subsequent refinement.

To remove clearly irrelevant proposals, VERA parses the referring expression into structured semantic cues using a query parsing function $\Psi(\cdot)$:

$$\mathcal{Z}_Q = \Psi(Q) = (c_Q, a_Q, r_Q), \tag{2}$$

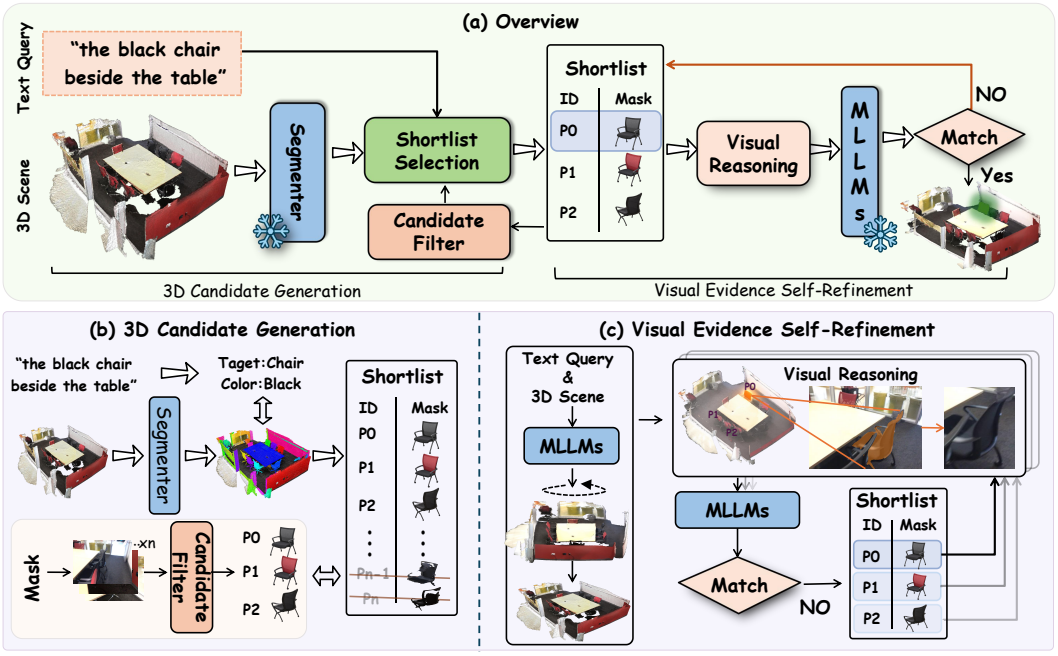

Figure 2: Overview of VERA. Given a 3D scene and a referring expression, 3D Candidate Generation first constructs a query-relevant candidate hypothesis space while preserving the corresponding point-level masks. Visual Evidence Self-Refinement then verifies and conditionally refines candidate hypotheses using multi-granularity visual evidence, and returns the selected instance mask according to the task formulation.

where $\mathcal{Z}_Q$ denotes the structured semantic representation of the referring expression, and c_Q, a_Q, and r_Q represent the target category, appearance attributes, and spatial relations, respectively. The target category is first used for coarse semantic screening:

$$\mathcal{R} = \{O_i \in \mathcal{O} \mid \phi(O_i, c_Q) = 1\}, \tag{3}$$

where $\phi(\cdot)$ denotes a rule-based category consistency function that checks whether the semantic label of proposal O_i matches the target category c_Q, and $\mathcal{R}$ represents the retained category-consistent candidate set. This step removes obviously irrelevant proposals while retaining plausible target candidates.

To improve target candidate recall, VERA introduces a Candidate Filter (CF) in the 3D Candidate Generation stage. Implementation details are provided in Appendix A.1. Specifically, Grounding DINO is invoked only when the initial candidate pool lacks sufficient category-consistent proposals. The detected 2D regions are associated with existing 3D proposals through 2D–3D projection consistency to recover potentially missed candidates while preserving their original point-level masks.

After semantic screening and optional recovery, the remaining candidates are pruned and ranked using a rule-based ranking function according to semantic relevance and geometric consistency, producing an ordered shortlist:

$$\widetilde{\mathcal{C}} = \mathrm{Rank}(\mathcal{C}) = (\tilde{O}_1, \tilde{O}_2, \ldots, \tilde{O}_K), \tag{4}$$

where $\widetilde{\mathcal{C}}$ denotes the ordered candidate shortlist, $\mathcal{C}$ contains candidates that are semantically relevant to the referring expression. $\tilde{O}_i$ represents the i-th ranked candidate instance, and K is the number of retained candidates. The resulting shortlist is used for subsequent visual evidence verification.

3.3 Visual Evidence Self-Refinement

To address the ambiguity among visually similar candidates, we propose the Visual Evidence Self-Refinement module. Instead of directly accepting the top-ranked candidate, Visual Evidence Self-Refinement verifies the candidate using multi-granularity visual evidence and revises the prediction when necessary.

3.3.1 MULTI-GRANULARITY VISUAL EVIDENCE CONSTRUCTION

To support this evidence-driven verification process, VERA constructs a multi-granularity visual evidence space that provides complementary visual observations for evaluating candidate hypotheses. The goal of evidence construction is not to learn additional visual features, but to enable the MLLMs to inspect candidate hypotheses from different perspectives through explicit visual evidence. Given a candidate instance $\tilde{O}_i \in \widetilde{\mathcal{C}}$, VERA constructs its associated visual evidence set $\mathcal{E}_i$:

$$\mathcal{E}_i = \{E^v, E_i^c, E_i^o\}, \tag{5}$$

Specifically, E^v, E_i^c, and E_i^o denote view-level, context-level, and object-level evidence, respectively. These three types of evidence correspond to different reasoning requirements in 3D referring expressions: view-level evidence captures global scene layout, context-level evidence models spatial relations between the candidate and surrounding instances, and object-level evidence provides fine-grained appearance cues for verifying object-specific attributes. The detailed prompt template is provided in Appendix A.2.

View-level evidence. To capture the global spatial layout of the scene, VERA renders the 3D scene from multiple bird's-eye-view (BEV) perspectives and selects the viewpoint that provides the most informative spatial evidence for the referring expression. Specifically, given multiple BEV observations $\{I_v^{bev}\}_{v=1}^{N_v}$, VERA prompts a frozen MLLM to select the most relevant viewpoint:

$$v^* = \mathcal{M}_{\text{MLLM}}\left(Q, \{I_v^{bev}\}_{v=1}^{N_v}\right), \qquad v^* \in \{1, \ldots, N_v\}, \tag{6}$$

where $\mathcal{M}_{\text{MLLM}}(\cdot)$ denotes the frozen multimodal large language model used for viewpoint selection, and N_v denotes the number of rendered BEV viewpoints. The MLLMs directly outputs a discrete viewpoint index rather than a continuous ranking score. The selected BEV observation is then used as the view-level evidence $E^v = I_{v^*}^{bev}$.

Context-level evidence. To preserve the spatial relationships between a candidate object and its surrounding scene, VERA constructs context-level evidence by selecting the most informative RGB observation according to the geometric visibility of the candidate object. Specifically, for each candidate object $\tilde{O}_i$, its 3D mask is projected onto registered RGB views, and the RGB view with the highest context relevance is selected:

$$k_i^* = \arg\max_k S_{\text{ctx}}\left(\tilde{O}_i, I_k\right), \tag{7}$$

where $S_{\text{ctx}}(\cdot)$ denotes the context relevance score for RGB view selection. The score considers the geometric visibility and image-space coverage of the projected candidate, including visible projected points, region coverage, depth visibility, and image center alignment.

The selected RGB frame together with the projected candidate location serves as the context-level evidence:

$$E_i^c = \left(I_{k_i^*}, \Pi_{k_i^*}(\tilde{O}_i)\right), \tag{8}$$

where $\Pi_{k_i^*}(\tilde{O}_i)$ denotes the projected location of candidate object $\tilde{O}_i$ on the selected RGB view. Since the context-level evidence preserves both the candidate object and its surrounding scene, it provides spatial information for verifying relational constraints in the referring expression.

Object-level evidence. Finally, to capture fine-grained visual information of each candidate object, VERA constructs object-level evidence based on the projected location of the candidate in the corresponding RGB observation. Specifically, for each candidate object $\tilde{O}_i$, VERA extracts a local object region from the selected RGB view using its 2D projected region:

$$E_i^o = \text{Crop}\left(I_{k_i^*}, \Pi_{k_i^*}(\tilde{O}_i)\right), \tag{9}$$

where $\mathrm{Crop}(\cdot)$ denotes extracting the local object region from the RGB observation according to the projected region, $I_{k_i^*}$ represents the RGB observation selected for candidate object $\tilde{O}_i$, and $\Pi_{k_i^*}(\tilde{O}_i)$ denotes the corresponding 2D projection of the candidate object on this view. The resulting object-level evidence focuses on the appearance and fine-grained visual details of the candidate object, providing complementary cues for verifying object categories and attributes. Together with view-level and context-level evidence, it supports the candidate verification process of VERA.

3.3.2 Evidence-based Candidate Verification

The candidate generation stage provides an initial ranking of candidate objects, but the top-ranked candidate may still be ambiguous in scenes containing multiple similar instances. Therefore, VERA verifies the candidates using the constructed multi-granularity visual evidence rather than directly accepting the highest-ranked result.

For a candidate object $\tilde{O}_i$, the MLLM predicts a binary verification result based on the referring expression Q and its corresponding visual evidence $\mathcal{E}_i$:

$$m_i = \mathcal{M}_{\mathrm{MLLM}}\left(Q, \mathcal{E}_i\right), \qquad m_i \in \{0, 1\}, \quad i = 1, \ldots, K, \tag{10}$$

where $m_i = 1$ indicates that candidate object $\tilde{O}_i$ is consistent with the referring expression under the provided visual evidence. VERA first verifies the highest-ranked candidate $\tilde{O}_t = \tilde{O}_1$. If $m_1 = 1$, the candidate is directly accepted. Otherwise, VERA examines the remaining candidates following their original ranking and selects the first verified candidate as a replacement candidate $\tilde{O}_c$. If no candidate receives positive verification, VERA retains the original prediction.

Since verification only indicates that a candidate is consistent with the referring expression and does not guarantee its superiority over the initial prediction, VERA further performs a pairwise comparison between $\tilde{O}_t$ and $\tilde{O}_c$ using their object-level evidence:

$$(w_{tc}, q_{tc}) = \mathcal{M}_{\mathrm{MLLM}}\left(Q, \mathrm{Concat}(E_t^o, E_c^o)\right), \tag{11}$$

where w_{tc} denotes the candidate preferred by the MLLMs and q_{tc} denotes the corresponding confidence. The replacement is performed only when the replacement candidate is explicitly preferred and satisfies $q_{tc} \geq \gamma$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate VERA on three widely used 3D referring benchmarks: ScanRefer (Chen et al., 2020), Multi3DRefer (Zhang et al., 2023), and Nr3D (Achlioptas et al., 2020), all built on ScanNet (Dai et al., 2017). ScanRefer contains 51,583 referring expressions over 800 indoor scenes and evaluates single-object localization, with queries grouped into Unique and Multiple subsets according to the presence of same-category distractors. Multi3DRefer extends this setting to a variable number of referred objects, including Zero Target, Single Target, and Multiple Targets cases, with 61,926 expressions covering 11,609 objects. Nr3D contains 41,503 expressions collected through a referential game and emphasizes fine-grained discrimination among same-category instances.

Evaluation metrics. We evaluate point-level 3D referring segmentation using Acc@0.25, Acc@0.5, and mIoU. Acc@t measures the percentage of queries whose predicted mask achieves an IoU of at least t with the ground-truth mask, while mIoU denotes the mean mask IoU over all queries. All metrics are reported as percentages. For the Candidate Filter ablation, we additionally report Shortlist Oracle, i.e., the percentage of queries for which the target object is included in the candidate shortlist, measuring candidate recall before final prediction. For Nr3D and Multi3DRefer, we follow the official evaluation protocols and splits provided by each benchmark. All reported results are computed using the corresponding standard evaluation metrics.

Implementation details. We use Qwen3-VL-32B-Instruct (Bai et al., 2025) as the default MLLMs and serve it locally with vLLM. Instance proposals are generated using a pretrained OneFormer3D

Table 1: Comparison on Nr3D and Multi3DRefer. All values are mask-based metrics (%).

Methods	Venue	Zero-shot	Unique			Multiple			Overall		
			Acc@0.25	Acc@0.5	mIoU	Acc@0.25	Acc@0.5	mIoU	Acc@0.25	Acc@0.5	mIoU
Expert methods											
3D-STMN	AAAI 2024	✗	89.3	84.0	74.5	46.2	29.2	31.1	54.6	39.8	39.5
RefMask3D	ACM MM 2024	✗	89.6	84.7	–	48.1	40.8	–	55.9	49.2	44.9
IPDN	AAAI 2025	✗	91.5	88.0	77.9	53.1	47.0	43.6	60.6	54.9	50.2
Position3D	CVPR 2026	✗	92.0	88.9	77.6	54.1	48.7	44.7	61.5	56.1	51.0
MVGGT	CVPR 2026	✗	83.6	74.5	65.2	49.2	33.5	33.8	55.9	41.5	39.9
LLM-based methods											
SegPoint	ECCV 2024	✗	–	–	–	–	–	–	–	–	41.7
Reason3D	3DV 2025	✗	88.4	84.2	74.6	50.5	31.7	34.1	57.9	41.9	42.0
3D-LLaVA	CVPR 2025	✗	–	–	–	–	–	–	–	–	43.3
MR-COSMO	AAAI 2026	✗	–	–	–	–	–	–	–	–	45.6
GenSplat	CVPR 2026	✗	–	–	–	–	–	–	61.3	42.4	43.6
VERA (Ours)	–	✓	87.6	83.9	74.7	48.4	44.8	41.0	55.6	52.0	47.3

Table 2: Comparison on Nr3D and Multi3DRefer. All values are mask-based metrics (%).

Methods	Venue	Zero-shot	Nr3D			Multi3DRefer		
			Acc@0.25	Acc@0.5	mIoU	Acc@0.25	Acc@0.5	mIoU
Expert methods								
TGNN	AAAI 2021	✗	25.7	20.9	19.1	–	–	–
3D-STMN	AAAI 2024	✗	41.5	25.8	27.6	60.4	40.9	43.0
RG-SAN	NeurIPS 2024	✗	40.1	28.7	28.6	–	–	–
MDIN	ACM MM 2024	✗	–	–	–	67.0	44.7	47.5
IPDN	AAAI 2025	✗	–	–	–	71.5	50.0	51.7
Position3D	CVPR 2026	✗	–	–	–	72.3	52.7	53.2
LLM-based methods								
SegPoint	ECCV 2024	✗	–	–	32.2	–	–	36.1
3D-LLaVA	CVPR 2025	✗	–	–	32.6	–	–	42.7
VERA (Ours)	–	✓	41.9	37.7	34.7	58.4	45.7	44.8

backbone (Kolodiazhnyi et al., 2024), followed by instance-level deduplication and gated candidate supplementation with Grounding DINO (Liu et al., 2024). For visual evidence, we render one top-down view and four oblique global views, from which the MLLMs selects a query-relevant viewpoint for subsequent verification together with candidate-specific RGB views and local object crops. The shortlist size is dynamically determined, and the MLLMs verifies at most 12 candidates during inference. All models are used off-the-shelf without task-specific fine-tuning.

For single-object referring benchmarks (ScanRefer and Nr3D), VERA follows the default selective refinement strategy and outputs the selected candidate mask. For Multi3DRefer, which allows zero, single, and multiple referred objects, VERA performs candidate-wise visual verification and aggregates all positively verified candidates as the final prediction set.

4.2 Quantitative Comparison

We compare VERA with existing approaches on ScanRefer, Nr3D, and Multi3DRefer. As shown in Table 1, VERA achieves competitive performance on ScanRefer without task-specific training, obtaining 52.0% Acc@0.5 and 47.3% mIoU overall. These results exceed the reported overall Acc@0.5 and mIoU of the listed LLM-based methods, while remaining below specialist methods such as IPDN and Position3D. On the Multiple split, VERA achieves 44.8% Acc@0.5 and 41.0% mIoU, outperforming Reason3D and 3D-STMN on both metrics. These results demonstrate competitive instance disambiguation in scenes containing same-category distractors. The contribution of visual evidence-based verification is further examined in the ablation study.

Table 2 extends the evaluation to Nr3D and Multi3DRefer. Without task-specific training, VERA achieves 41.9% Acc@0.25, 37.7% Acc@0.5, and 34.7% mIoU on Nr3D, exceeding the reported results of the listed reference methods. On Multi3DRefer, VERA obtains 58.41% Acc@0.25, 45.73% Acc@0.5, and 44.8% mIoU. Its mIoU surpasses SegPoint and 3D-LLaVA, although a gap remains to IPDN and Position3D. Together, these results support VERA's applicability to single-object and

Table 3: VERA ablations on ScanRefer, Nr3D, and Multi3DRefer. All results are mask-based mIoU (%).

Variant	ScanRefer			Nr3D	M3DRef
	Unique	Multiple	All	mIoU	mIoU
Effect of Visual Evidence Self-Refinement					
w/o Self-Refinement	70.7	29.8	37.4	30.3	34.2
Visual Evidence Ablation					
w/o Panorama	73.9	40.2	46.5	33.6	43.7
w/o RGB	72.9	36.5	43.3	31.0	35.5
w/o Crop	73.4	38.0	44.5	32.2	38.9
VERA (Full)	74.7	41.0	47.3	34.7	44.8

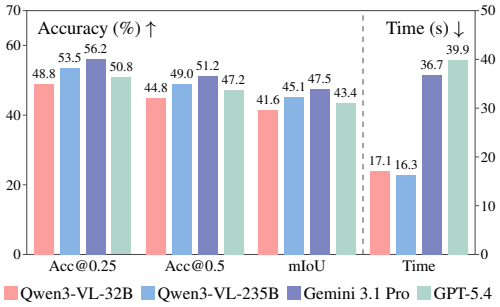

Figure 3: VLM backbone comparison: accuracy (%, left axis; higher is better) and time (s/query, right axis; lower is better).

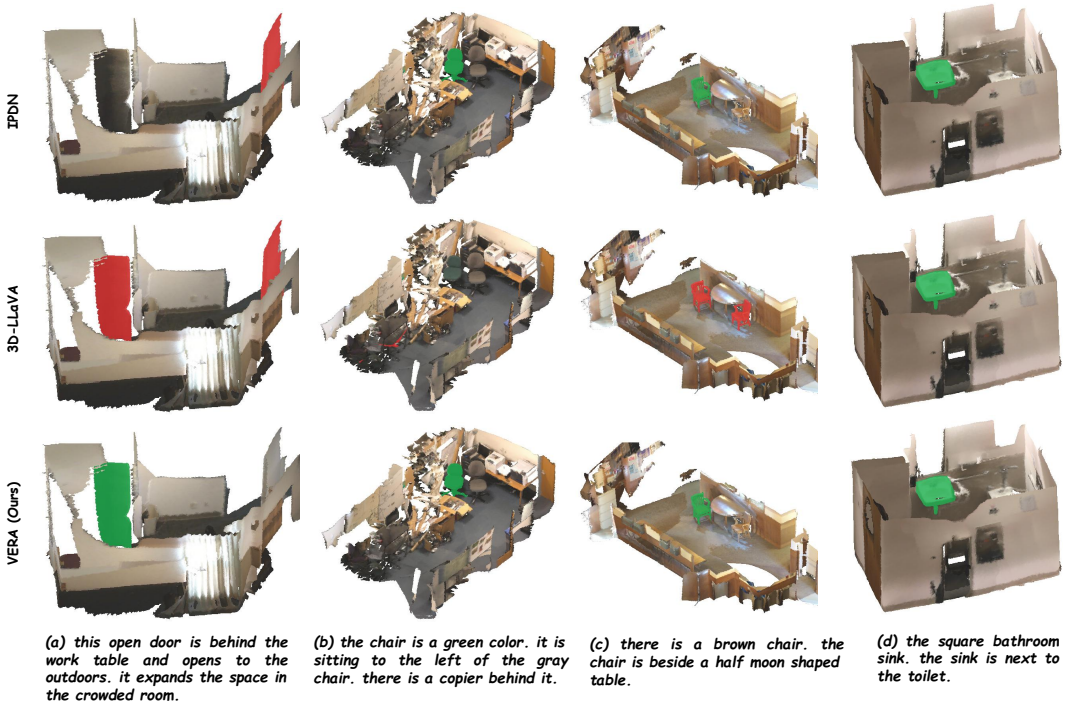

Figure 4: Qualitative comparison of 3D referring segmentation results. Red and green regions indicate incorrect and correct predictions, respectively. VERA effectively resolves ambiguous instances through visual evidence-based candidate verification.

zero-/multi-target referring segmentation settings without task-specific training. As all three benchmarks are built on ScanNet, the findings concern different referring settings within the same underlying scene dataset.

4.3 Qualitative Comparison

Effect of MLLMs backbone selection. To investigate the dependence of VERA on the underlying MLLMs, we evaluate four off-the-shelf backbones using the same candidate generation, visual evidence construction, and prompts. We randomly sample 400 queries from the ScanRefer validation split with a fixed seed and report referring segmentation performance and inference time in Figure 3. VERA remains effective with different backbones without additional training. Gemini 3.1 Pro achieves the highest referring segmentation performance among the evaluated backbones. Compared with Qwen3-VL-32B, Qwen3-VL-235B improves the performance by 4.7, 6.4, and 5.6 points in Acc@0.25, Acc@0.5, and mIoU, respectively. These results demonstrate that VERA is compatible with different MLLMs backbones and that backbone choice affects referring segmentation quality and inference efficiency.

Table 4: Effect of Candidate Filter (CF) on candidate quality. Values in parentheses indicate improvements over w/o CF.

Variant	Oracle@0.25	Oracle@0.50	Oracle mIoU	Mean $\|\mathcal{C}\|$	CF Invocation Rate
ScanRefer					
w/o CF	85.0	80.5	72.9	8.1	–
w/ CF	**86.9** (+1.9)	**82.2** (+1.7)	**74.4** (+1.5)	8.6 (+0.4)	23.1%
Nr3D					
w/o CF	77.2	71.3	64.9	6.3	–
w/ CF	**80.2** (+3.0)	**73.8** (+2.5)	**67.9** (+2.3)	6.7 (+0.3)	16.8%
Multi3DRefer					
w/o CF	82.2	67.3	62.8	7.5	–
w/ CF	**84.2** (+2.1)	**69.2** (+1.8)	**64.4** (+1.6)	8.0 (+0.5)	28.8%

Qualitative results. Figure 4 presents qualitative comparisons between VERA and existing methods on challenging 3D referring segmentation examples. Existing approaches may fail when multiple instances share similar appearances or spatial configurations, as selecting a plausible instance does not guarantee the correct point-level mask. In contrast, VERA verifies candidate hypotheses through Visual Evidence Self-Refinement by jointly considering object appearance, local context, and scene-level spatial relationships. This evidence-driven verification enables VERA to resolve ambiguous instances and preserve the correspondence between the selected candidate and its 3D mask, avoiding incorrect instance selection under similar semantic conditions.

4.4 Ablation Study

We conduct ablation studies to evaluate Visual Evidence Self-Refinement, multi-granularity visual evidence, and Candidate Filter (CF).

Effect of visual evidence self-refinement. As shown in Table 3, removing Visual Evidence Self-Refinement reduces mIoU across all three benchmarks. On ScanRefer, the full module improves mIoU by 11.2 percentage points on the Multiple split (29.8% to 41.0%), compared with 4.0 points on the Unique split (70.7% to 74.7%). This larger gain on the Multiple split supports the effectiveness of visual verification in distinguishing same-category instances beyond the initial candidate ranking.

Effect of different visual evidence sources. Removing any individual evidence source reduces performance across all three benchmarks, supporting their complementary contributions. Removing RGB contextual views causes the largest overall mIoU drop on each benchmark, particularly on Multi3DRefer (44.8% to 35.5%). This highlights the importance of candidate-specific RGB observations, which provide both appearance and contextual cues. Object-level crops and panoramic views offer additional benefits.

Effect of candidate filter. As shown in Table 4, CF improves candidate coverage by recovering potentially missed proposals through open-vocabulary detection. The gains in oracle performance indicate a more complete candidate set before verification, while the small increase in candidate count suggests that this improved coverage requires only a limited expansion of the candidate space.

5 Conclusion

In this work, we present VERA, a training-free framework for 3D referring segmentation that leverages pretrained MLLMs and off-the-shelf perception tools without task-specific supervision. VERA formulates 3D-RES as an evidence-driven hypothesis verification process, where candidate instances are progressively verified through multi-granularity visual evidence while maintaining their associated 3D masks. Extensive experiments on standard 3D referring segmentation benchmarks demonstrate the effectiveness of VERA and highlight the potential of pretrained multimodal models for training-free 3D understanding.

AI USE STATEMENT

Following the conference guidelines regarding generative artificial intelligence models (GenAI), we declare that GenAI tools were used only for manuscript preparation, including language polishing and proofreading. All research ideas, framework designs, algorithmic developments, experimental settings, result analyses, and scientific conclusions were independently developed and verified by the authors.

ETHICS STATEMENT

This work investigates training-free 3D referring segmentation based on pretrained multimodal and 3D perception models. We conduct experiments on publicly available benchmark datasets collected from existing 3D scene understanding tasks. Our method does not involve human subjects, personal data collection, or user interaction. We acknowledge that the pretrained models and datasets used in this work may inherit biases or limitations from their original sources; however, our method does not introduce additional ethical concerns beyond those associated with the original models and datasets.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed framework, implementation details, evaluation protocols, and prompt templates in the supplementary material. Our method is a training-free framework that relies on publicly available datasets and frozen pretrained models without task-specific training or additional parameter optimization. We will release the implementation and necessary resources to reproduce our results upon publication.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pp. 422–440. Springer, 2020.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19107–19117. IEEE, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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*, pp. 202–221. Springer, 2020.

Junlong Chen, Jens Grubert, and Per Ola Kristensson. Analyzing multimodal interaction strategies for llm-assisted manipulation of 3d scenes. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 206–216. IEEE, 2025a.

Junlong Chen, Jens Grubert, and Per Ola Kristensson. A comparative study of speech-and-pointing and disocclusion mini-map techniques for object selection in virtual reality. In *Proceedings of the 2025 ACM Symposium on Spatial User Interaction*, pp. 1–12, 2025b.

Qi Chen, Changli Wu, Jiayi Ji, Yiwei Ma, Danni Yang, and Xiaoshuai Sun. Ipdn: Image-enhanced prompt decoding network for 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2132–2140, 2025c.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26418–26428. IEEE, 2024.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38689–38699, 2026.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. Llmr: Real-time prompting of interactive worlds using large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2024.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3772–3782. IEEE, 2025.

Jiayu Ding, Meilu Song, Xiaoyi Zhang, Hongbo Jin, Yichen Jin, and Xiangtian Si. Zerosplat: Generalized referring segmentation in 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 147–163. Springer, 2026.

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, 2026.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. In *International Conference on Learning Representations*, volume 2025, pp. 13354–13385, 2025.

Shuting He and Henghui Ding. Refmask3d: Language-guided transformer for 3d referring segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8316–8325, 2024.

Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pp. 349–367. Springer, 2024.

Jiale Huang and Shangfei Wang. Saqn: Semantic-based adaptive query network for 3d referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38627–38636, 2026.

Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. Mllm-for3d: Adapting multimodal large language model for 3d reasoning segmentation. *Advances in Neural Information Processing Systems*, 38:54961–54984, 2026a.

Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. In *2025 International Conference on 3D Vision (3DV)*, pp. 1177–1186. IEEE, 2025a.

Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 1610–1618, 2021.

Sheng-Yu Huang, Jaesung Choe, Yu-Chiang Frank Wang, and Cheng Sun. Openvoxel: Training-free grouping and captioning voxels for open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2601.09575*, 2026b.

Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv preprint arXiv:2507.23478*, 2025b.

Haichao Jiang, Tianming Liang, Wei-Shi Zheng, and Jian-Fang Hu. Refer-agent: A collaborative multi-agent system with reasoning and reflection for referring video object segmentation. *arXiv preprint arXiv:2602.03595*, 2026.

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. *Advances in Neural Information Processing Systems*, 38:165549–165576, 2026.

Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20943–20953. IEEE, 2024.

Chade Li, Pengju Zhang, and Yihong Wu. Mr-cosmo: Visual-text memory recall and direct cross-modal alignment method for query-driven 3d segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6010–6018, 2026.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3707–3717. IEEE, 2025.

Fang Liu, Yuhao Liu, Ke Xu, Gerhard Petrus Hancke, and Rynson WH Lau. Gensplat: bridging the generalization gap in 3dgs language comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5221–5231, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

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*, volume 2026, pp. 157277–157302, 2026.

Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4551–4559, 2024.

Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29469–29478. IEEE, 2025.

Duotun Wang, Linjie Qiu, Boyu Li, Qianxi Liu, Xiaoying Wei, Jianhao Chen, Zeyu Wang, and Mingming Fan. Focalselect: Improving occluded objects acquisition with heuristic selection and disambiguation in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 31(5):3285–3295, 2025.

Wenxin Wang, Bo Zhang, Feng Chen, Zixuan Wang, Wen Li, Changsheng Li, and Yinjie Lei. Agentic collaborative cognition for zero-shot 3d understanding. In *European Conference on Computer Vision*, pp. 371–390. Springer, 2026a.

Yabing Wang, Zhuotao Tian, Le Wang, Zheng Qin, and Sanping Zhou. Spatial matters: Position-guided 3d referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39486–39496, 2026b.

Changli Wu, Qi Chen, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. Rg-san: Rule-guided spatial awareness network for end-to-end 3d referring expression segmentation. *Advances in Neural Information Processing Systems*, 37:110972–110999, 2024a.

Changli Wu, Yihang Liu, Jiayi Ji, Yiwei Ma, Haowei Wang, Gen Luo, Henghui Ding, Xiaoshuai Sun, and Rongrong Ji. 3d-gres: Generalized 3d referring expression segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7852–7861, 2024b.

Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5940–5948, 2024c.

Changli Wu, Haodong Wang, Jiayi Ji, Yutian Yao, Chunsai Du, Jihua Kang, Yanwei Fu, and Liujuan Cao. Mvggt: Multimodal visual geometry grounded transformer for multiview 3d referring expression segmentation. *arXiv preprint arXiv:2601.06874*, 2026a.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 2026b.

Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024a.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pp. 131–147. Springer, 2024b.

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. IEEE, 2024.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14203–14214. IEEE, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026.

Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, Tong He, and Yanyong Zhang. Agent3d-zero: An agent for zero-shot 3d understanding. In *European Conference on Computer Vision*, pp. 186–202. Springer, 2024.

Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15179–15179. IEEE, 2023.

A APPENDIX

In the Appendix, we provide the following additional materials:

- implementation details of VERA in Appendix A.1;
- prompt templates used by the MLLM in Appendix A.2;

A.1 IMPLEMENTATION DETAILS

We provide additional implementation details of VERA to facilitate reproducibility.

Given a 3D scene, VERA first obtains instance proposals from OneFormer3D. Each proposal is associated with a point-level mask, which is preserved throughout the verification process. The referring expression is decomposed into target category, appearance attributes, and spatial relations. Candidate proposals are ranked using a training-free heuristic score that combines category agreement, appearance consistency, spatial relation matching, proposal confidence, and object size:

$$s_i = 2.0s_i^{cat} + 0.7s_i^{col} + 1.0s_i^{spa} + 0.4c_i + 0.05z_i,$$

where s_i^{cat}, s_i^{col}, and s_i^{spa} denote category, appearance, and spatial matching scores, respectively. c_i and z_i represent the normalized proposal confidence and object size cues. The coefficients are fixed across all experiments and are not optimized using benchmark annotations. The ranked candidates form a dynamic shortlist for subsequent visual verification.

For visual evidence construction, VERA generates complementary observations at different granularities. Global scene evidence is obtained by rendering one top-down view and four oblique views of the 3D scene. The VLM selects the query-relevant viewpoint from these observations. For candidate-specific evidence, each candidate mask is projected onto registered ScanNet RGB frames. The frame with the most informative candidate observation is selected according to the depth-visible projection quality, considering the visible projected points, projected region coverage, and image-center preference. The selected frame provides contextual evidence through the projected candidate region and object-level evidence through a tight crop of the candidate.

For inference, we use Qwen3-VL-32B-Instruct in an off-the-shelf manner and serve the model locally with vLLM. All VLM calls use greedy decoding ($\mathrm{temperature} = 0$) and require structured JSON outputs for verification. The three visual evidence images (global view, RGB context, and object crop) are concatenated into a single visual input.

For single-object benchmarks (ScanRefer and Nr3D), VERA follows the selective refinement strategy. It first verifies the initial hypothesis and examines alternative candidates only when the current hypothesis lacks sufficient visual support. The initial hypothesis is replaced only when the alternative candidate is preferred through direct comparison. For Multi3DRefer, VERA independently verifies each candidate and aggregates all positively verified candidates as the final prediction set.

A.2 PROMPT TEMPLATES

We provide the prompt templates used by VERA for different visual reasoning stages. The prompts are fixed for all experiments and are not optimized using benchmark annotations. Specifically, Table 5 presents the prompt used for viewpoint selection, where the VLM identifies the most informative global view according to the referring expression. Table 6 shows the candidate verification prompt, which is used to determine whether a shortlisted candidate matches the query based on multi-granularity visual evidence. Table 7 provides the optional pairwise comparison prompt used for selective refinement, where the VLM compares the initial hypothesis and a challenger candidate when replacement is required.

Table 5: Prompt template for viewpoint selection.

Prompt Template for Viewpoint Selection

You are good at finding the object in a 3D scene based on a given query description. These images show different views of a room. You need to find the {target_class} in this query description: "{text}"

The image is a grid of five panoramic renderings. The view number is in the top-left of each panel: 1 top (bird's-eye) | 2 left | 3 right | 4 up | 5 down.

Please review all view images to find the target object and select the view that you can see the target object most clearly. Do not pick an object ID yet. Only pick the viewpoint.

Output your answer in JSON format with these keys:

```
{
  "reasoning": "Explain how you identified the target object,
               and why you choose this view.",
  "view": "2"
}
```

Table 6: Prompt template for visual grounding of one shortlist candidate.

Prompt Template for Visual Grounding

You ground a referring expression to ONE indoor object from visual evidence. JSON only.

You see THREE images of ONE object. Image 1: a panoramic 3D rendering of the room; this object's ID is marked. Image 2: a ScanNet RGB camera frame with the same object boxed. Image 3: a close-up crop of that box.

This is the {CURRENT HYPOTHESIS | ALTERNATIVE candidate} {cid}. Decide whether it answers the query. Referring query: {query}

Check the image(s) against the query yourself (category, appearance, neighbors, spatial language). Do not assume this ID is correct or incorrect. match=true only if THIS object is a valid answer to the query.

For Multi3DRefer, the query may refer to zero, one, or several objects; match=true iff THIS object itself is among the referred targets.

Return JSON:

```
{
  "id": "{cid}",
  "match": true,
  "score": 0.0,
  "fail": [],
  "why": "short"
}
```

Table 7: Prompt template for the optional pairwise switch.

Prompt Template for Pairwise Comparison

You fairly compare two boxed ScanNet RGB crops. Wrong-class or landmark-as-target must lose. Do not default to the current tip.

Fair pairwise: choose which candidate better matches the referring expression. NO default preference for A or B. Each panel is the boxed object only.

OBJECT TO FIND: {`target_category`} Query: {`description`}
A (LEFT) = {`id_A`} (current tip) B (RIGHT) = {`id_B`} (challenger)

1) First filter by TARGET CATEGORY. If A is the wrong class or a query LANDMARK and B matches the target, `winner=B`.
2) Then use color / spatial cues to break same-class ties.
3) `winner=B` when B is clearly the better referred instance.
4) `winner=A` when A is clearly better OR the evidence is truly ambiguous.

Return JSON only:

```
{
  "winner": "A|B",
  "confidence": 0.0,
  "why": "decisive cue"
}
```