Walking Through 3D Spaces: Spatial Routing for Referring 3D Gaussian Splatting Segmentation

Peilin Ji[1] **Changshuo Wang**[2] **Yong Zhang**[1,*] **Shuting He**[1,*]
[1]Shanghai University of Finance and Economics [2]University College London
peilin.ji806@gmail.com, {zhang.yong, shuting.he}@sufe.edu.cn

Abstract

This paper introduces Walk3D, a novel 3D Gaussian Splatting (3DGS) framework explicitly designed for 3D referring segmentation, particularly excelling at interpreting complex, long-sentence descriptions. Our primary motivation stems from observing that existing 3DGS-based open-vocabulary methods mainly focus on simple, category-level object segmentation. These methods struggle to precisely segment targets described by highly detailed, multi-sentence queries due to weak compositional language understanding and a lack of structural scene awareness. To tackle these dense and complex linguistic constraints, we propose a spatial routing network that decodes these queries into sequential chain-of-thought instructions upon a scene graph. Specifically, we construct a static supernode graph based on the geometric and semantic properties of the 3DGS space, where a spatial Graph Neural Network (GNN) dynamically pathfinds using the decomposed linguistic chain to accurately segment the referred target. Extensive experiments on complex 3D referring segmentation benchmarks demonstrate the superiority of our proposed method in interpreting lengthy, multi-conditional queries. The code is publicly available at https://github.com/SPT806/Walk3D.

1 Introduction

3D Gaussian Splatting (3DGS) [17] provides an explicit, point-based representation that enables efficient differentiable rendering and a direct linkage between 2D observations and 3D structures, offering a unique advantage over implicit methods like NeRF [29]. This architectural foundation facilitates the perception and interaction with 3D environments via natural language, which underpins next-generation autonomous navigation [32, 20], immersive AR/VR [25, 15, 37], and generative XR content editing [43, 2].

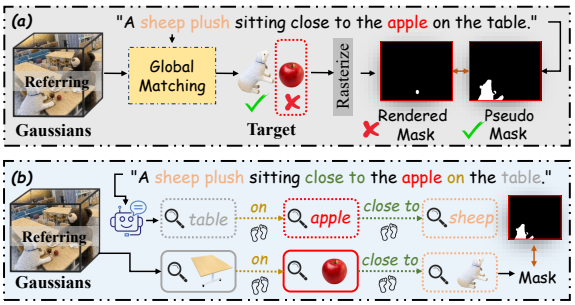


Figure 1: Comparison of R3DGS methods. (a) Prior methods rely on flat, global matching, failing on complex queries. (b) Walk3D leverages an LLM to parse queries into sequential spatial chains, enabling dynamic, step-by-step routing for precise segmentation.

Building on these strengths, recent research has moved beyond simple novel-view synthesis toward 3D semantic scene understanding [49, 33, 39]. While pioneering methods such as LangSplat [33], OpenGaussian [44], and Feature-3DGS [49] have demonstrated effective word-level open-vocabulary perception, they mostly rely on fixed-pattern descriptors and

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

global feature matching. Consequently, these approaches often struggle when faced with complex, long-form spatial descriptions or intricate relational queries that require explicit geometric reasoning. To bridge this semantic-geometric gap, the field has recently pivoted toward Referring 3D Gaussian Splatting Segmentation (R3DGS) [11]. Unlike generic perception, R3DGS requires instance-specific localization driven by complex linguistic expressions, necessitating a profound understanding of scenes where targets may be occluded or visually ambiguous. Early efforts like ReferSplat [11] utilized language-aware fields and confidence-weighted pseudo-masks for alignment. Subsequent advancements, such as OpenVoxel [14] and CaRF [40], have achieved further improvements.

Despite this progress, a fundamental challenge remains: the persistent disconnect between high-level linguistic reasoning and low-level geometric primitives. As shown in Fig. 1, current implicit matching paradigms often lead to fragmented segmentation in cluttered environments, as they lack the structural depth to resolve multi-hop logic. We attribute these persistent limitations to two main factors: **1) Insufficient structural reasoning depth:** Current paradigms lack the necessary structural depth to resolve complex multi-hop logic or localized spatial anchors. This often results in fragmented segmentation performance, particularly in cluttered and occluded 3D environments. **2) Missing explicit positional priors:** The absence of explicit spatial grounding within standard language encoders makes it difficult to disambiguate target objects through implicit semantic correlations alone. Consequently, these models struggle to establish accurate fine-grained mappings between high-level reasoning and low-level geometric primitives.

To address these limitations, we present Walk3D, a framework that handles the R3DGS task by employing a dynamic, step-wise routing process instead of conventional static matching. Drawing inspiration from human spatial cognition, Walk3D introduces a Chain-of-Thought (CoT) routing mechanism grounded on a learned 3D topological scene graph. By clustering raw Gaussian primitives into semantically cohesive supernodes, Walk3D decomposes complex queries into a sequence of executable spatial operations (e.g., *find anchor* $\rightarrow$ *move relation* $\rightarrow$ *find target*). This logic-driven traversal allows the model to progressively propagate semantic information and prune irrelevant distractors, converging on the target with high precision.

Extensive experiments on three benchmarks (Ref-LERF [11], LERF-OVS [33], and 3D-OVS [25]) demonstrate that Walk3D achieves state-of-the-art results, reaching mIoU scores of 40.0%, 56.2%, and 94.2%, respectively. Our main contributions are: 1) **Walk3D**, a two-stage framework addressing the R3DGS task by employing dynamic spatial pathfinding instead of conventional static matching; 2) a **spatial routing network** leveraging supernode graphs and multi-modal edge-gating to translate complex linguistic constraints into step-wise CoT navigation; 3) **comprehensive evaluations** establishing new SOTA results and demonstrating robust performance in dense scenes with multi-conditional queries.

2 Related Work

3D Neural Representations. The field of 3D representation has witnessed a significant paradigm shift from implicit modeling to explicit structures. Neural Radiance Fields (NeRF)[29] pioneered this wave by utilizing continuous implicit neural networks for high-quality novel view synthesis. However, its implicit nature inherently limits direct spatial manipulation and leads to extended optimization durations. While methods like Instant-NGP[31] introduced hash encoding for acceleration, they still operate within the implicit volume rendering framework. More recently, 3D Gaussian Splatting (3DGS)[17] revolutionized the field by proposing a purely explicit and discrete paradigm, representing continuous scenes using a collection of anisotropic 3D Gaussian distributions. In this work, we adopt 3DGS strictly as our foundational 3D representation. Rather than treating it merely as a rendering pipeline, we leverage its explicit geometric properties—where each Gaussian serves as a discrete, addressable spatial entity—to construct a topology-aware scene graph, bridging the gap between 3D neural representations and complex structural reasoning.

3D Segmentation in Gaussian Splatting. 3D Gaussian Splatting (3DGS) utilizes 2D foundation models such as CLIP [35], SAM [19], and GroundingDINO [26] for cross-modal and spatial priors. Current research in 3D semantic [21] and open-vocabulary segmentation [33, 44, 49, 45] follows three main trajectories: feature distillation [33, 49, 39, 8, 10], mask lifting [28, 13, 48, 27], and feed-forward construction [1, 16, 22, 38]. Despite these advancements, existing methods often fail to handle the complex spatial reasoning required for long-form linguistic queries.

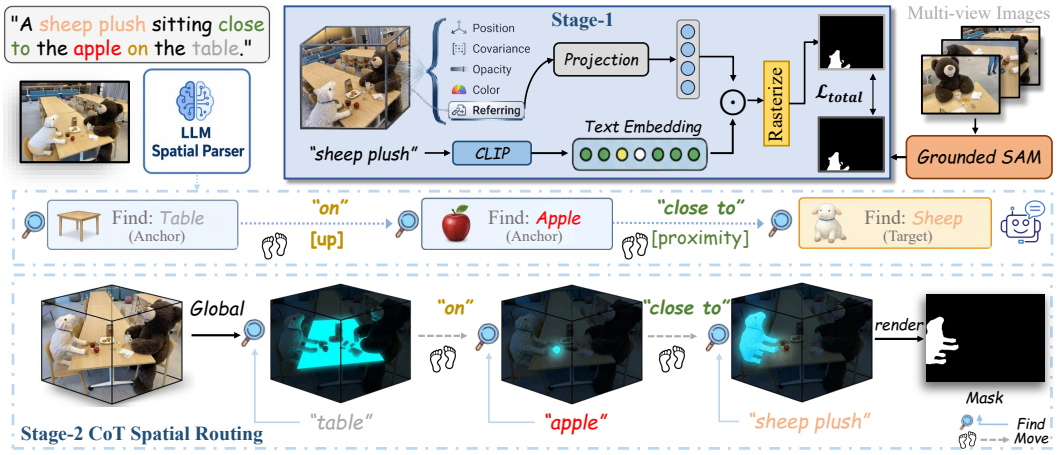


Figure 2: **Overview of the Walk3D Pipeline.** An LLM first translates complex queries into a decoupled Spatial CoT. Concurrently, Stage-1 aligns 3D Gaussians with 2D VLMs to establish semantic priors. Guided by these logical and semantic cues, Stage-2 executes step-by-step spatial routing to precisely segment the target.

Referring 3D Gaussian Splatting Segmentation. R3DGS [11, 40, 41, 14] extends 2D referring segmentation [47, 3–5, 7, 6, 12] to 3D scenes using diverse alignment strategies. ReferSplat [11] achieves global semantic alignment by integrating language features into the 3DGS rendering pipeline via confidence-weighted pseudo-masks. Conversely, OpenVoxel [14] provides a training-free alternative, employing MLLM-captioned voxel groups for direct text-to-text retrieval, effectively bypassing traditional CLIP-based embedding alignment. Existing approaches struggle to resolve relative spatial dependencies, as their reliance on global sentence-level matching fails to bind specific descriptors to corresponding targets. When dealing with long, complex descriptions, matching the sentence as a whole often leads to incorrect associations between objects and their spatial relations. This global approach fails to capture the step-by-step reasoning needed to resolve multi-link spatial queries. To overcome these limitations, Walk3D abstracts the continuous 3D field into a static, topologically-aware supernode graph. By executing step-wise Chain-of-Thought (CoT) spatial routing that strictly decouples entity grounding (*"find"*) from relational pathfinding (*"move"*), Walk3D preserves multi-step spatial logic, ensuring highly precise and geometrically coherent 3D grounding.

3 Method

3.1 Overview

In this section, we introduce **Walk3D**, a framework that addresses the R3DGS task through explicit structural reasoning instead of conventional static matching. As shown in Fig. 2, an LLM first parses expressions into a Spatial Chain-of-Thought (Sec. 3.2), followed by Stage-1 Semantic Grounding (Sec. 3.3) to initialize Gaussians with language-aware features. The scene is then abstracted into a topologically-aware supernode graph (Sec. 3.4). As illustrated in Fig. 3, Stage-2 (Sec. 3.5) executes step-wise Spatial Routing on this graph to resolve complex relational constraints. Finally, a zero-compute hard assignment maps these reasoning scores back to dense Gaussians for high-fidelity 2D mask rendering.

3.2 LLM-Driven Spatial Chain-of-Thought Generation

Human spatial instructions are inherently compositional and often feature reference inversions (e.g., "A is near B" requires finding B first). To systematically parse these without relying on brittle NLP pipelines, we leverage LLM to abstract raw text into a strict Spatial Chain-of-Thought (CoT).

We formulate the referring expression as a directed traversal graph, alternating between node and edge operations: $\mathcal{P} = \{(type_1, text_1), \ldots, (type_L, text_L)\}$, where $type \in \{$"*find*", "*move*"$\}$. For

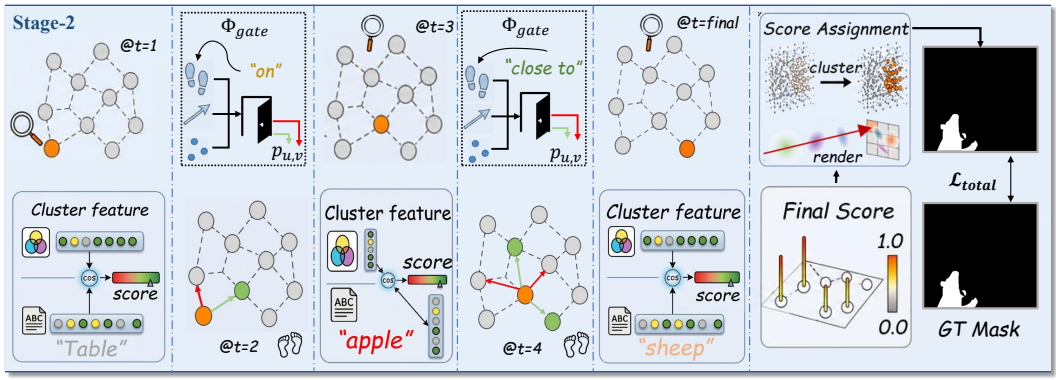


Figure 3: **Step-wise Spatial Routing in Stage-2.** The network alternates between entity grounding ("find", e.g., @t=1, 3) via feature similarity, and relational traversal ("move", e.g., @t=2, 4) via a topological gate Φ_{gate}. Final scores are mapped to Gaussians for rendering and supervision.

instance, *"A is near B"* is parsed as:

$$\mathcal{P} = \{(\textit{find}, B), (\textit{move}, \text{“}\textit{near}\text{”}_{[\text{proximity}]}), (\textit{find}, A)\}, \tag{1}$$

here, the *type* acts as an explicit execution signal for the downstream GNN. A *find* operation dictates semantic node grounding to isolate relevant supernodes. Conversely, a *move* operation triggers spatial edge traversal, with its $text$ providing a standardized topological prior to gate message passing. To mitigate LLM hallucinations and ensure deterministic executability, we enforce two protocols: **(1) Pre-Reasoning & Target Anchoring:** The LLM explicitly extracts background anchors and the main target prior to CoT generation, grounding the traversal to prevent semantic drift. **(2) Topological Standardization:** Open-vocabulary prepositions are mapped to a closed set of fixed tags (e.g., [up], [inside]) to resolve spatial ambiguity and yield deterministic edge operations. Ultimately, this strategy transforms linguistic ambiguity into strictly executable graph traversal paths, with Sec. 4 further validating our framework's broad applicability by analyzing the impact of various 7B-9B scale LLM backbones on CoT quality and final performance.

3.3 Stage-1: Base Semantic Grounding and Alignment

Before performing complex spatial reasoning, the raw 3D scene must acquire foundational semantic awareness. We extend standard 3DGS by augmenting each Gaussian g_i with a low-dimensional, learnable referring feature vector $f_{r,i} \in \mathbb{R}^d$ to construct a dense 3D Gaussian Referring Field.

To map these compact representations into the rich multi-modal space, we employ a non-linear projection network, yielding the aligned Gaussian feature $\tilde{f}_{r,i}$. We extract the textual embedding f_{text} using CLIP. To capture fine-grained semantic features and overcome the extreme foreground-background imbalance, we introduce a learnable temperature scalar κ. The similarity score S_i is dynamically amplified to inherently sharpen the feature boundaries:

$$S_i = \sigma\left((\tilde{f}_{r,i} \cdot f_{text}) \cdot e^{\kappa}\right). \tag{2}$$

Following the paradigm introduced by ReferSplat [11], we render the point-wise Gaussian similarity scores directly into a 2D mask $\hat{M}$ via 3DGS alpha-compositing. To train these Stage-1 semantic features without 3D annotations, we optimize the rendered mask $\hat{M}$ against 2D pseudo ground-truth masks M_{gt} extracted from the training views, supervised by our robust boundary-aware loss (detailed in Sec. 3.6).

3.4 Topologically-Aware Supernode Graph Construction

Running Graph Neural Networks directly on millions of dense Gaussians is computationally intractable and topologically noisy. To establish a stable physical skeleton for spatial reasoning, we abstract the continuous scene into a static Supernode Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We apply Mini-Batch

K-Means clustering on a tri-modal descriptor concatenating normalized physical coordinates μ_i, color attributes c_i, and the semantic features $f_{r,i}$. This joint clustering ensures that spatially adjacent points with similar appearances but distinct semantics are correctly partitioned. The feature H_k for each supernode is initialized by average-pooling the referring features of its constituent Gaussians. Next, we construct the directed edge set $\mathcal{E}$ using K-Nearest Neighbors based on spatial centers P_k. Each edge $e_{i,j}$ is assigned a 4-dimensional geometric attribute vector $E_{i,j}$ that concatenates the L_2 distance and the normalized 3D direction vector:

$$E_{i,j} = \left[||P_j - P_i||_2, \frac{P_j - P_i}{||P_j - P_i||_2 + \epsilon} \right]. \tag{3}$$

3.5 Stage-2: CoT Spatial Routing

Guided by the LLM-generated CoT chain $\mathcal{P}$, Walk3D executes step-wise spatial routing on the supernode graph. We maintain a dynamic activation state $A^{(t)} \in [0,1]^K$ for all K supernodes. This state explicitly tracks the probability distribution of the target region at each reasoning step t. Depending on the operation $type$, the network sequentially updates the graph state:

Node Grounding (*type == "find"*):
The objective of this operation is to semantically anchor specific text entities within the 3D graph. For each supernode k, we evaluate its semantic alignment with the current step's text embedding $f_{text}^{(t)}$ by computing the cosine similarity against the supernode feature H_k, amplified by a learnable temperature scalar e^{κ}. To trace the reasoning sequence, the node's activation state $A_k^{(t)}$ is updated as a progressive logical intersection:

$$A_k^{(t)} = \begin{cases} \sigma\left((H_k \cdot f_{text}^{(t)}) \cdot e^{\kappa} \right), & t = 1, \\ \min\left(A_k^{(t-1)}, \sigma\left((H_k \cdot f_{text}^{(t)}) \cdot e^{\kappa} \right) \right), & t > 1, \end{cases} \tag{4}$$

here, $A_k^{(t-1)}$ represents the historical confidence accumulated from previous steps. By enforcing a strict minimum bottleneck, this formulation ensures that a supernode retains a high activation state if and only if it matches the current semantic query and continuously satisfies the historical trajectory of the spatial chain. Such a mechanism decisively filters out geometric noise and prevents the model from hallucinating targets based on partial matches.

Edge-Gated Traversal (*type == "move"*):
This operation performs spatial hops, shifting the activation from previously grounded anchors to new candidate regions guided by topological text (e.g., "[topological_inside]"). To govern these spatial transitions, we employ a MLP to learn the gating mapping Φ_{gate} that maps the joint multi-modal context into a directed passage probability $p_{u,v} \in [0,1]$ for the edge $u \to v$:

$$p_{u,v} = \Phi_{gate}\left(H_u, H_v, E_{u,v}, f_{text}^{(t)} \right). \tag{5}$$

This MLP dynamically evaluates whether the transition from the source semantics H_u to the destination semantics H_v logically aligns with the directional text prior $f_{text}^{(t)}$, rigorously conditioned on the underlying physical edge geometry $E_{u,v}$. Once the passage probabilities are resolved, the activation is propagated along the directed graph:

$$A_v^{(t)} = \text{clamp}\left(\sum_{u \in \mathcal{N}(v)} A_u^{(t-1)} \cdot p_{u,v},\ 0,\ 1.0 \right). \tag{6}$$

We aggregate the inbound confidence $A_u^{(t-1)}$ from neighboring nodes $u \in \mathcal{N}(v)$, weighted by the learned transfer probability $p_{u,v}$. The clamp$(\cdot, 0, 1.0)$ operation serves as a robust boundary constraint, preventing numerical overflow during recursive message passing while ensuring the activation states maintain a valid probabilistic interpretation within the unit interval $[0,1]$. By bounding the cumulative signal, this mechanism facilitates stable convergence across complex spatial trajectories and ensures that the final reasoning scores can be directly mapped to high-fidelity masks.

By alternating *find* and *move* operations, the network navigates the 3D scene graph, shifting activation from anchors to the final target at the CoT chain's terminus. After the final CoT step, the ultimate

supernode scores $A^{final} \in \mathbb{R}^K$ are directly assigned to their constituent 3D Gaussians. Following ReferSplat [11], we utilize Grounded SAM[36] to extract pseudo masks, which serve as supervision for both Stage 1 and Stage 2 training. These point-wise scores are subsequently rendered into the final 2D mask via standard 3DGS alpha-compositing.

3.6 Robust Optimization Objectives

To explicitly address the severe class imbalance and boundary ambiguity inherent in 3D referring segmentation, both stages are supervised end-to-end using a compound loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{focal} + \lambda_{dice}\mathcal{L}_{dice}. \tag{7}$$

Focal Loss for Class Imbalance: To mitigate the extreme class imbalance inherent in 3D spatial grounding, we employ a modified Focal Loss:

$$\mathcal{L}_{\text{focal}} = \frac{1}{N}\sum_{i=1}^{N} \alpha_t \cdot \omega \cdot \text{BCE}(p, y), \quad \text{where } \omega = \begin{cases} (1-p_t)^{1.0}, & y > 0.5, \\ (1-p_t)^{\gamma}, & y \leq 0.5. \end{cases} \tag{8}$$

We adopt a class-dependent modulating factor ω instead of a uniform hyperparameter [23]. For background ($y \leq 0.5$), we set $\gamma = 2.0$ to suppress noise and false positives. For positive targets ($y > 0.5$), a linear penalty is employed to prevent gradient attenuation and encourage decisive activations. Additionally, the balancing factor α_t is set to 0.7 for foreground targets to prioritize the recovery of small-scale entities.

Dice Loss for Boundary Precision: To enforce sharp geometric boundaries and ensure that activated Gaussians align closely with physical object contours, we incorporate Dice Loss. Defining $\epsilon = 10^{-5}$ as a smoothing term for numerical stability, the objective is formulated as:

$$\mathcal{L}_{dice} = 1 - \frac{2\sum(p \cdot y) + \epsilon}{\sum p + \sum y + \epsilon}. \tag{9}$$

In summary, this compound loss function provides robust optimization guarantees for achieving precise spatial grounding in severely imbalanced 3D environments by synergistically suppressing background noise, sharpening geometric boundaries, and penalizing global spurious activations.

4 Experiments

4.1 Experimental Setup

Datasets. Our empirical validation spans three representative benchmarks, each probing different aspects of 3D segmentation: Ref-LERF [11] is utilized for instance-level referring understanding; LERF-OVS [33] is employed to assess open-vocabulary 3DGS scaling across diverse scenes; and 3D-OVS [25] is used to test large-scale spatial grounding with category-level and room-level semantics.

Metrics. In line with established protocols [11, 40, 41], the mean Intersection-over-Union (mIoU) serves as our standard evaluation metric. For any specific query-view pair, the overlap between the rendered prediction mask $\hat{Y}$ and the corresponding ground-truth Y is quantified as

$$\text{IoU}(Y, \hat{Y}) = \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|}. \tag{10}$$

Implementation Details. All experiments are implemented in PyTorch on a single NVIDIA L20 GPU. We extract language features using CLIP (ViT-B/16) [35] and generate CoT spatial reasoning paths via the DeepSeek-V3 API [24]. 3D Gaussian features are initialized from pre-trained 3DGS models [17], where the reference features are denoted as $f_{r,i} \in \mathbb{R}^d$ with $d = 256$. The network is optimized via Adam [18] in two distinct stages (2 epochs for Stage-1, 3 for Stage-2). We apply a differential learning rate: 1×10^{-3} for Gaussian reference features and 1×10^{-4} for the spatial router parameters. To supervise precise segmentation, we set $\lambda_{\text{dice}} = 0.5$. All other hyperparameters and configurations strictly follow ReferSplat [11].

Table 1: Comparison with state-of-the-art methods on the Ref-LERF dataset in terms of mIoU (↑).

Method	Venue	Ramen	Figurines	Teatime	Kitchen	Average
SPIn-NeRF[30]	[ICCV'23]	7.3	9.7	11.7	10.3	9.8
Grounded SAM[36]	[ArXiv'24]	14.1	16.0	16.9	16.2	15.8
LangSplat[33]	[CVPR'24]	12.0	17.9	7.6	17.9	13.9
GS-Grouping[48]	[ECCV'24]	27.9	8.6	14.8	6.3	14.4
GOI[34]	[MM'24]	27.1	16.5	22.9	15.7	20.6
ReferSplat[11]	[ICML'25]	35.2	25.7	31.3	24.4	29.2
CaRF[40]	[ICME'26]	33.5	28.7	29.7	24.7	29.2
Walk3D (Ours)	–	**50.0**	**37.9**	**36.0**	**36.1**	**40.0**
Improvement	–	+16.5	+9.2	+6.3	+11.4	+10.8

Table 2: 3D open-vocabulary segmentation results on the LERF-OVS dataset. We report the mIoU (%).

Method	Venue	Ram.	Fig.	Tea.	Kit.
Feature-3DGS[49]	[CVPR'24]	43.7	58.8	40.5	39.6
LEGaussians[39]	[CVPR'24]	46.0	60.3	40.8	39.4
LangSplat[33]	[CVPR'24]	51.2	65.1	44.7	44.5
GS-Grouping[48]	[[ECCV'24]]	45.5	60.9	40.0	38.7
GOI[34]	[MM'24]	52.6	63.7	44.5	41.4
ReferSplat[11]	[ICML'25]	53.1	64.1	50.1	43.3
Walk3D(Ours)	–	**58.4**	**66.5**	**50.5**	**49.2**

Table 3: 3D open-vocabulary segmentation results on the 3D-OVS dataset. mIoU (%).

Method	bed	ben.	roo.	sof.	law.
Feature-3DGS[49]	83.5	90.7	84.7	86.9	93.4
LEGaussians[39]	84.9	91.1	86.0	87.8	92.5
LangSplat[33]	92.5	94.2	94.1	90.0	96.1
GS-Grouping[48]	83.0	91.5	85.9	87.3	90.6
GOI[34]	89.4	92.8	91.3	85.6	94.1
ReferSplat[11]	90.2	93.8	94.1	90.8	95.5
Walk3D(Ours)	**93.0**	**94.8**	**94.5**	**91.9**	**96.7**

4.2 Comparisons with State-of-the-Arts

Referring 3DGS Segmentation. Walk3D achieves state-of-the-art (SOTA) results on Ref-LERF (Tab. 1), delivering an average mIoU of 40.0% and outperforming the strongest baselines, ReferSplat [11] and CaRF [40], by 10.8% mIoU (40.0% vs. 29.2%). The performance advantage is most evident in cluttered scenes like Ramen (+16.5% mIoU) and Kitchen (+11.4% mIoU), where "flat-matching" models like ReferSplat struggle due to a lack of spatial logic. Unlike the static one-step alignment in existing methods, Walk3D employs a Chain-of-Thought (CoT) strategy over scene topology, replacing brittle matching with a dynamic, path-finding process. This structured approach ensures robust, geometrically coherent grounding in complex environments, leading SOTA by a wide margin.

3D Open-vocabulary Segmentation. To assess generalizability, we evaluate Walk3D on LERF-OVS[33] and 3D-OVS[25] benchmarks (Tab. 2, 3). Despite its focus on multi-step reasoning, Walk3D sets new SOTAs with 56.2% and 94.2% average mIoU, respectively. This superiority in category-level tasks stems from our two-stage design: Stage-1(Sec. 3.3) establishes high-fidelity semantic priors that provide a faithful foundation for both general scene parsing and the subsequent Stage-2(Sec. 3.5) spatial routing. This confirms that our structured representation scales effectively from complex reasoning to broad, room-scale scene understanding.

4.3 Ablation Studies

We ablate Walk3D's core components to validate its design. Tab. 5 confirms the joint necessity of Stage-1 priors and Stage-2 routing. Further studies validate Stage-2's CoT reasoning and routing mechanism (Tab. 6), the necessity of LLM parsing (Tab. 4), and architectural choices (Tab. 7). Ultimately, Walk3D successfully bridges 3D semantic grounding with topological navigation, resolving the fundamental limitations of R3DGS.

Table 4: Effect of LLM parsing on CoT.

Method (LLM Backbone)	Ram.	Kit.
ReferSplat[11](w/o CoT)	35.2	24.4
Walk3D (LLaMA3.1-8B[9])	42.6	30.2
Walk3D (Qwen2.5-7B[46])	41.7	29.5
Walk3D (Gemma2-9B[42])	43.4	31.1

Effect of LLM Parsing on CoT. As shown in Tab. 4, we investigate the impact of language model selection on Chain-of-Thought decomposition and overall segmentation quality. Compared to the baseline ReferSplat without CoT decomposition, equipping Walk3D with LLM-guided spatial

Table 5: Ablation on training paradigms. We evaluate Semantic Prior (Stage 1) and Spatial Router (Stage 2) on Ref-LERF.

Index	Semantic Prior	Spatial Router	Teatime (mIoU)	Figurines (mIoU)
Baseline	✗	✗	24.6	22.3
Setting 1	✓	✗	25.8	23.4
Setting 2	✗	✓	20.1	18.7
Walk3D	✓	✓	**36.0** +11.4	**37.9** +15.6

Table 6: Ablation of Stage 2 components. We evaluate the individual and combined effects of CoT reasoning and the Spatial Router.

Setting	Teatime (mIoU)	Figurines (mIoU)
w/o CoT & Spatial Router	25.8	23.4
w/o Spatial Router	29.5	28.7
w/o CoT	31.2	30.5
w/ CoT & Spatial Router(Ours)	**36.0** +10.2	**37.9** +14.5

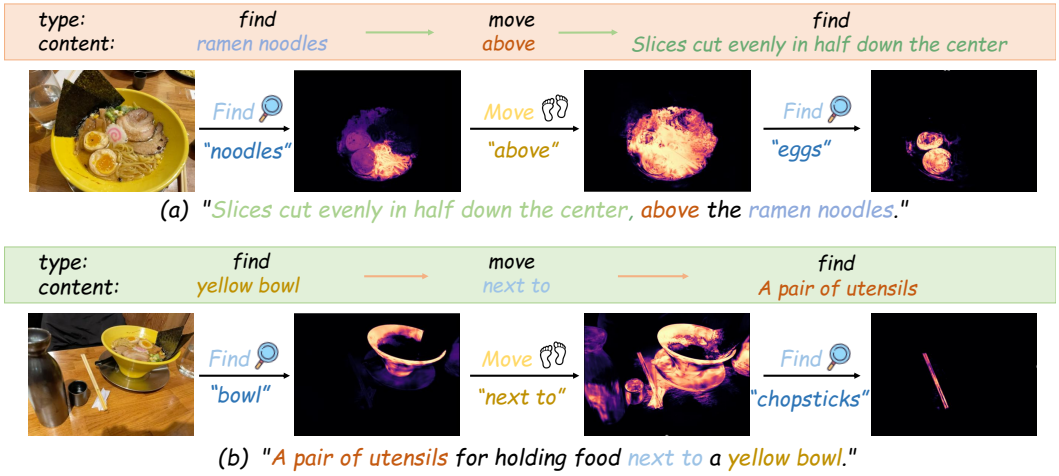


Figure 4: **Visualization of CoT routing.** Walk3D resolves complex queries via sequential spatial anchors (colors denote entities and relations). Activation heatmaps progressively sharpen as multi-step routing converges on the target.

routing consistently yields substantial performance gains across both scenes. Furthermore, Walk3D demonstrates remarkable stability across diverse open-source backbones—including LLaMA3.1-8B[9], Qwen2.5-7B[46], and Gemma2-9B[42]—confirming that our topological routing framework is robust and does not rely on the idiosyncrasies of a specific language model.

Effectiveness of Two-Stage Paradigm and Routing. Tab. 5 investigates module synergy against a global matching Baseline similar to ReferSplat[11]. Using only the Stage-1 semantic prior (Setting 1) yields marginal gains (25.8 vs. 24.6 mIoU in Teatime), failing to resolve spatial ambiguity. Conversely, using only the Stage-2 router (Setting 2) propagates errors without accurate anchors, causing a severe drop (20.1 mIoU in Teatime). Further dissecting Stage 2 (Tab. 6), we isolate its core mechanisms. The *w/o Spatial Router* setting retains structured step-by-step logic but disables topological graph propagation, whereas the *w/o CoT* setting forces multi-hop graph traversal guided only by a static, entangled global prompt. Ultimately, our full Walk3D framework proves that decoupling reasoning steps and executing them via explicit physical graph routing are jointly indispensable for success.

Ablation of Architectural Designs. We further investigate the impact of graph configuration, gating mechanisms, and optimization objectives in Tab. 7:

(a) Graph Resolution. We evaluate varying graph resolutions to mitigate semantic aliasing—where a single node conflates distinct entities—and establish an optimal topological configuration. As illustrated in Fig. 5, a coarse graph resolution with too few supernodes induces severe semantic aliasing (Fig. 5b), collapsing adjacent object boundaries and diluting discriminating features. Con-

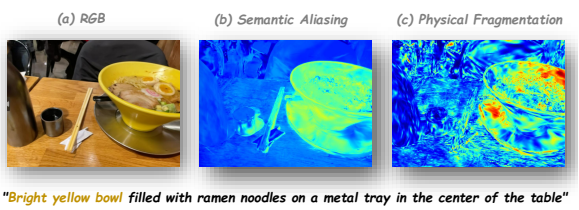


Figure 5: **Effect of graph resolution.** (a) RGB image. (b) Coarse graphs cause semantic aliasing. (c) Overly fine graphs cause physical fragmentation.

Table 7: Ablation study on architectural designs of Walk3D. Bold values indicate best performance.

Aspect	Setting	Teatime	Figurines	Avg.
(a) Graph Resolution	$K, E = 256, 4$ (Semantic Aliasing)	20.1	22.0	21.1
	$K, E = 512, 4$	29.4	31.9	30.7
	$K, E = 1024, 6$ **(Optimal Balance)**	**36.0**	**37.9**	**37.0**
	$K, E = 2048, 6$ (Physical Fragmentation)	34.7	35.3	35.0
(b) Gating Design	w/o $f_{text}^{(t)}$	33.9	34.7	34.3
	w/o $H_u + H_v$	35.4	36.2	35.8
	w/ $H_u + H_v + E_{u,v} + f_{text}^{(t)}$**(Ours)(see eq. 5)**	**36.0**	**37.9**	**37.0**
(c) Optimization Loss	Focal	32.9	33.4	33.2
	Focal + Dice(Ours)(see eq. 7)	**36.0**	**37.9**	**37.0**

versely, an excessively fine-grained resolution causes physical fragmentation (Fig. 5c), breaking cohesive objects into disjoint sub-clusters and disrupting holistic topological navigation. Balancing these trade-offs is thus essential for robust spatial reasoning.

(b) Gating Design. We validate our feature integration strategy (Eq. 5), demonstrating the necessity of comprehensively fusing geometric topology and textual semantics rather than relying on either cue independently. As shown in Tab. 7, removing dynamic language guidance causes a marked performance drop, as topological routing loses explicit directional guidance and diffuses aimlessly across neighbors. Likewise, omitting node representations weakens entity grounding. The full design confirms that dynamically conditioning edge transitions on both local topological context and linguistic intent is crucial for accurate multi-hop routing.

(c) Optimization Loss. We confirm the effectiveness of utilizing a composite Focal and Dice objective (Eq. 7) over a baseline relying solely on a standard Focal loss. As reported in Tab. 7, optimizing with Focal loss alone struggles to handle extreme foreground-background class imbalance, often leading to fragmented masks or over-smoothed predictions on fine-grained targets. Integrating the Dice loss enforces global shape consistency and volumetric overlap, ensuring sharp object boundaries and robust convergence for complex 3D referring segmentation.

4.4 Qualitative Analysis

Visualizing the Chain-of-Thought Routing. Fig. 4 visualizes Walk3D's interpretable routing, which decomposes complex queries into sequential *find* and *move* operations. By anchoring on spatial references (e.g., "ramen noodles" or "yellow bowl") to navigate topological relations ("above", "next to"), Walk3D prunes distractors to pinpoint fine-grained targets. This CoT process transforms R3DGS from black-box association into an explainable trajectory of logical transitions, effectively mimicking human-like spatial reasoning.

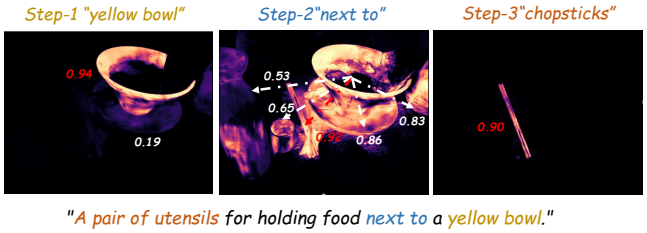


Figure 6: **Step-wise node activations.** Heatmaps illustrate activation scores across sequential reasoning steps, tracking the routing trajectory from anchor to target.

Specifically, as depicted in Fig. 6, intermediate node activations trace this multi-hop progression: Step 1 localizes the anchor reference, Step 2 propagates attention along relational edges while pruning distractors, and Step 3 concentrates activation on the target, validating our robust topological routing.

Robustness in Failure Scenarios. Fig. 7 illustrates Walk3D's error tolerance: **(a) Ambiguity.** Faced with visually similar entities, Walk3D maintains spatial awareness and relational cues where baselines fail due to referential ambiguity. **(b) CoT Errors.** Despite LLM hallucinations producing imprecise operations (e.g., "scene"), Walk3D roams via **scattered graph nodes** to reach the target's vicinity, ensuring robust coarse grounding despite reasoning flaws.

Qualitative Comparison on Ref-LERF. Fig. 8 visualizes qualitative results on Ref-LERF[11] across four scenes, comparing RGB inputs, ReferSplat[11], Ours method, and Ground Truth (GT). While ReferSplat yields fragmented masks due to spatial ambiguity, Walk3D leverages topological

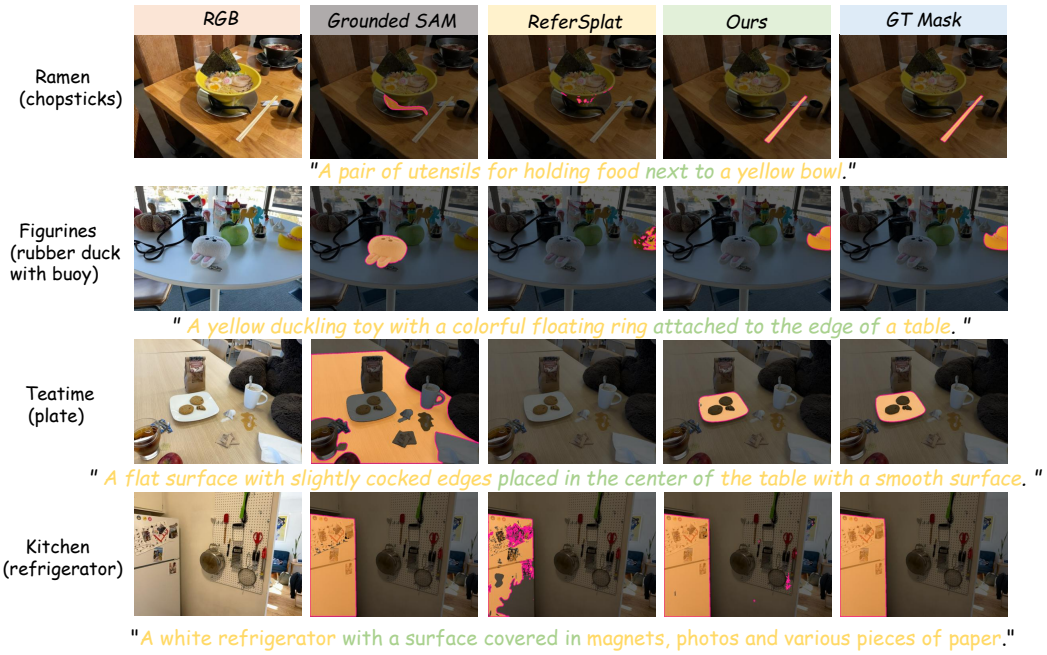


Figure 8: Qualitative results on Ref-LERF. In the linguistic prompts, yellow text denotes target entities, while green text indicates spatial relationships and positional descriptions.

routing to prune distractors, producing precise, Gaussian-level segmentations that closely align with the GT.

5 Conclusion

We presented **Walk3D**, a 3DGS-GNN framework bridging semantic priors and topological navigation for referring 3D Gaussian Splatting segmentation. By mapping complex queries to explicit reasoning paths over a supernode graph, Walk3D transforms ambiguous instructions into interpretable trajectories, achieving state-of-the-art performance and effectively suppressing distractors in cluttered scenes. Extensive ablations further confirm the indispensable synergy between Stage-1 dense priors and Stage-2 topological routing in resolving complex spatial ambiguities.



Figure 7: **Failure analysis. (a).** Object ambiguity causes incorrect target selection. **(b).** CoT errors lead to faulty routing.

Limitation. While Walk3D effectively resolves multi-hop spatial queries, several limitations remain. Routing accuracy stays tied to the fidelity of LLM reasoning chains, where noisy steps may induce path deviations. Furthermore, scaling to expansive environments incurs non-trivial graph computation overhead. Future efforts will target error-resilient self-correction and lightweight scaling for real-time robotics.

Acknowledgments and Disclosure of Funding

This project was sponsored by Natural Science Foundation of Shanghai 25ZR1402138 and 11ZR1412100, Shanghai Pujiang Programme 24PJD030, and IQuest Research. Changshuo Wang

was supported in part by the European Union's Horizon 2024 Research and Innovation Programme for the Marie Skłodowska-Curie Actions under Grant No. 101211118.

References

[1] Kangjie Chen, BingQuan Dai, Minghan Qin, Dongbin Zhang, Peihao Li, Yingshuang Zou, and Haoqian Wang. Slgaussian: Fast language gaussian splatting in sparse views. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3047–3056, 2025.

[2] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21476–21485, 2024.

[3] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023.

[4] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. Mevis: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[5] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. Mosev2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025.

[6] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Yu-Gang Jiang. Grex: Generalized referring expression segmentation, comprehension, and generation. *International Journal of Computer Vision*, 134 (2):79, 2026.

[7] Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *International Journal of Computer Vision*, 134(9):405, 2026.

[8] Zihan Gao, Lingling Li, Licheng Jiao, Fang Liu, Xu Liu, Wenping Ma, Yuwei Guo, and Shuyuan Yang. Fast and efficient: Mask neural fields for 3d scene segmentation. *arXiv preprint arXiv:2407.01220*, 2024.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[11] Shuting He, Guangquan Jie, Changshuo Wang, Yun Zhou, Shuming Hu, Guanbin Li, and Henghui Ding. ReferSplat: Referring segmentation in 3d gaussian splatting. In *International Conference on Machine Learning*, 2025.

[12] Shuting He, Peilin Ji, Yitong Yang, Changshuo Wang, Jiayi Ji, Yinglin Wang, and Henghui Ding. A survey on 3d gaussian splatting applications: Segmentation, editing, and generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[13] Wenhao Hu, Wenhao Chai, Shengyu Hao, Xiaotong Cui, Xuexiang Wen, Jenq-Neng Hwang, and Gaoang Wang. Pointmap association and piecewise-plane constraint for consistent and compact 3d gaussian segmentation field. *arXiv preprint arXiv:2502.16303*, 2025.

[14] 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*, 2026.

[15] Xincheng Huang, Dieter Frehlich, Ziyi Xia, Peyman Gholami, and Robert Xiao. Demonstrating gaussian-nexus: Room-scale real-time ar/vr telepresence with gaussian splatting. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2026.

[16] Kim Jun-Seong, GeonU Kim, Kim Yu-Ji, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14137–14146, 2025.

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[20] Hailong Li, Kai Huang, Yuanhao Sun, Xiaohui Lei, Quanchun Yuan, Jinqi Zhang, and Xiaolan Lv. An autonomous navigation method for orchard mobile robots based on octree 3d point cloud optimization. *Frontiers in Plant Science*, 15:1510683, 2025.

[21] Hao Li, Minghan Qin, Zhengyu Zou, Diqi He, Xinhao Ji, Bohan Li, Bingquan Dai, Dingewn Zhang, and Junwei Han. Langsurf: Language-embedded surface gaussians for 3d scene understanding. *arXiv preprint arXiv:2412.17635*, 2024.

[22] Qijing Li, Jingxiang Sun, Liang An, Zhaoqi Su, Hongwen Zhang, and Yebin Liu. Semanticsplat: Feed-forward 3d scene understanding with language-aware gaussian fields. *arXiv preprint arXiv:2506.09565*, 2025.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2980–2988, 2017.

[24] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[25] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. In *Advances in Neural Information Processing Systems*, volume 36, pages 53433–53456, 2023.

[26] 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*, pages 38–55, 2024.

[27] Yiren Lu, Yunlai Zhou, Yiran Qiao, Chaoda Song, Tuo Liang, Jing Ma, and Yu Yin. Segment then splat: A unified approach for 3d open-vocabulary segmentation based on gaussian splatting. In *Advances in Neural Information Processing Systems*, 2025.

[28] Weijie Lyu, Xueting Li, Abhijit Kundu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Gaga: Group any gaussians via 3d-aware memory bank. *arXiv preprint arXiv:2404.07977*, 2024.

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.

[30] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023.

[31] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022.

[32] Shivam K Panda, Yongkyu Lee, and M Khalid Jawed. Agronav: Autonomous navigation framework for agricultural robots and vehicles using semantic segmentation and semantic line detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6272–6281, 2023.

[33] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024.

[34] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5328–5337, 2024.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[36] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[37] Yulin Shen, Boyu Li, Jiayang Huang, David Yip, and Zeyu Wang. Gaussianshopvr: Facilitating immersive 3d authoring using gaussian splatting in vr. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14, 2025.

[38] Yu Sheng, Jiajun Deng, Xinran Zhang, Yu Zhang, Bei Hua, Yanyong Zhang, and Jianmin Ji. Spatialsplat: Efficient semantic 3d from sparse unposed images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26404–26414, 2025.

[39] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024.

[40] Yuwen Tao, Kanglei Zhou, Xin Tan, and Yuan Xie. Carf: Enhancing multi-view consistency in referring 3d gaussian splatting segmentation. *arXiv preprint arXiv:2511.03992*, 2025.

[41] Yuwen Tao, Kanglei Zhou, Chang Li, and Liyuan Wang. Geometry-aware cross-modal graph alignment for referring segmentation in 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[42] Gemma Team. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL https://www.kaggle.com/m/3301.

[43] Minghao Wen, Shengjie Wu, Kangkan Wang, and Dong Liang. Intergsedit: Interactive 3d gaussian splatting editing with 3d geometry-consistent attention prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26136–26145, 2025.

[44] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In *Advances in Neural Information Processing Systems*, volume 37, pages 19114–19138, 2024.

[45] Yaxu Xie, Abdalla Arafa, Alireza Javanmardi, Christen Millerdurai, Jia Cheng Hu, Shaoxiang Wang, Alain Pagani, and Didier Stricker. Relags: Relational language gaussian splatting. In *CVPR*, 2026.

[46] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[47] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[48] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179, 2024.

[49] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024.

Appendix

A Preliminaries and Background

To make this paper self-contained, we briefly review the fundamental formulations utilized in our pipeline, specifically focusing on the rendering process and the acquisition of 2D pseudo masks that supervise our 3D referring fields.

A.1 Volumetric Rendering in 3DGS and ReferSplat

In this section, we briefly review the standard image rendering process in 3D Gaussian Splatting (3DGS) and its semantic extension introduced by ReferSplat, which serves as the foundational representation for our work.

In standard 3DGS, the color $C(v)$ at a given pixel v on the 2D image plane is computed by blending the contributions of N sorted overlapping Gaussians from front to back. This point-based α-blending process is formulated as:

$$C(v) = \sum_{i=1}^{N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{11}$$

where c_i is the color vector of the i-th Gaussian, and α_i is its evaluated opacity.

To address the R3DGS task, ReferSplat extends this rasterization process to establish language-perceptive fields. Specifically, instead of rendering RGB values, ReferSplat associates each 3D Gaussian with a scalar relevance score $m_i \in \mathbb{R}$, which represents its response to the input language query. The response on the 2D image plane at pixel v, denoted as $M(v)$, serves as the rendered segmentation mask and is computed exactly as:

$$M(v) = \sum_{i=1}^{N} m_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j). \tag{12}$$

In **Walk3D**, we directly adopt this score-based rasterization process from ReferSplat to obtain continuous referring fields, which provide robust 3D spatial grounding for our subsequent supernode graph construction and topological routing.

A.2 Pseudo Mask Generation

To supervise the 3D language-aligned fields during training, we acquire 2D pseudo masks from multi-view images following the robust candidate selection strategy proposed by ReferSplat.

Given a 2D image and a referring language expression, Grounded SAM is initially employed to generate a set of K candidate masks, denoted as $\{\hat{M}_k\}_{k=1}^{K}$. Each candidate mask is associated with a confidence score γ_k that exceeds a predefined threshold.

As noted in ReferSplat, relying solely on the highest confidence score or pure Intersection over Union (IoU) consistency often yields sub-optimal selections. To address this, a confidence-weighted IoU strategy is utilized to evaluate the overlapping score p_k for each candidate mask:

$$p_k = \sum_{j=1}^{K} \frac{\gamma_k \hat{M}_k \cap \gamma_j \hat{M}_j}{\gamma_k \hat{M}_k \cup \gamma_j \hat{M}_j}, \tag{13}$$

where γ_k and γ_j are the confidence scores corresponding to the candidate masks $\hat{M}_k$ and $\hat{M}_j$, respectively. This formulation effectively prioritizes masks that exhibit both high spatial consistency across candidates and high prediction confidence.

Finally, the candidate mask yielding the maximum score p_k is selected as the final pseudo mask $\hat{M}$. In **Walk3D**, we utilize these high-quality 2D pseudo masks as the ground truth to supervise the optimization of the 3D Gaussian relevance scores m_i.

B Additional Visualization Results

In this section, we provide more qualitative results across the four evaluation scenes from the Ref-LERF dataset to further demonstrate the effectiveness of Walk3D in diverse and complex indoor environments.

As shown in Figure B.1, we visualize the predicted segmentation masks and the corresponding heatmaps for diverse complex queries across the benchmark scenes. Specifically, we use a rendering matching score of 0.5

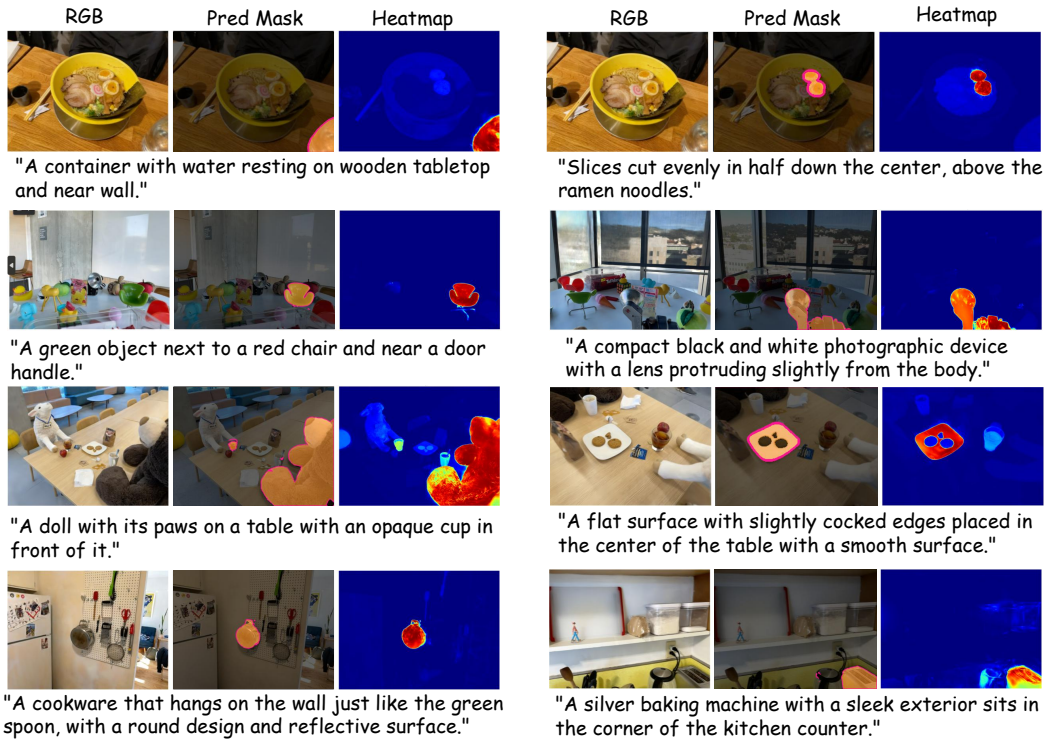


Figure B.1: Additional qualitative results on the Ref-LERF dataset. For each scene, we provide the **original RGB image**, the **predicted segmentation mask**, and the corresponding **heatmap**. These results demonstrate the fine-grained grounding and spatial reasoning capabilities of Walk3D across various complex queries.

as the threshold to obtain the masks. These results highlight several key strengths: (i) **Complex Relational Reasoning**, as our model successfully interprets intricate spatial relations (e.g., "next to," "above," and "near") to accurately localize target objects within cluttered environments; (ii) **Fine-grained Semantic Grounding**, where the language-aligned fields precisely capture small-scale and thin objects such as slices and hanging cookware, as evidenced by the high-resolution predicted masks and heatmaps; and (iii) **Open-Vocabulary Robustness**, demonstrating that our model's grounding remains effective and consistent across a wide range of diverse object categories and descriptive attributes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and Sec. 1

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 5

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 4

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: We will make the code publicly available once the paper is accepted.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4.1

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the high computational costs and resource constraints, it was not feasible to perform multiple runs for all experiments to report error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4.1

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully checked the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Sec. 5

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We politely cited the existing assets and read their usage license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Not applicable

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: Not applicable

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We employ an LLM as a core component of our methodology, with comprehensive details provided in Sec. 3.2

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.