Sparse3DPR: Training-Free 3D Hierarchical Scene Parsing and Task-Adaptive Subgraph Reasoning from Sparse RGB Views

Haida Feng[1,2*], **Hao Wei**[1*], **Zewen Xu**[1,2], **Haolin Wang**[1,2], **Chade Li**[1,2], **Yihong Wu**[1,2†]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{fenghaida2024@ia, weihao2019@ia, xuzewen2020@ia, wanghaolin2023@ia, lichade2021@ia, yhwu@nlpr.ia}.ac.cn

Abstract

Recently, large language models (LLMs) have been explored widely for 3D scene understanding. Among them, training-free approaches are gaining attention for their flexibility and generalization over training-based methods. However, they typically struggle with accuracy and efficiency in practical deployment. To address the problems, we propose Sparse3DPR, a novel training-free framework for open-ended scene understanding, which leverages the reasoning capabilities of pre-trained LLMs and requires only sparse-view RGB inputs. Specifically, we introduce a hierarchical plane-enhanced scene graph that supports open vocabulary and adopts dominant planar structures as spatial anchors, which enables clearer reasoning chains and more reliable high-level inferences. Furthermore, we design a task-adaptive subgraph extraction method to filter query-irrelevant information dynamically, reducing contextual noise and improving 3D scene reasoning efficiency and accuracy. Experimental results demonstrate the superiority of Sparse3DPR, which achieves a 28.7% EM@1 improvement and a 78.2% speedup compared with ConceptGraphs on the Space3D-Bench. Moreover, Sparse3DPR obtains comparable performance to training-based methods on ScanQA, with additional real-world experiments confirming its robustness and generalization capability.

Extended version — http://arxiv.org/abs/2511.07813

Introduction

Three-dimensional (3D) scene understanding is essential for embodied artificial intelligence, as it enables robots to understand, reason about, and execute natural language instructions within complex physical environments (Zhi et al. 2025). With the rapid advances of large language models (LLMs) (Liu et al. 2023; Achiam et al. 2023), particularly their strong capabilities in communication, commonsense reasoning, and open-world knowledge integration, have motivated LLM-based solutions for 3D scene understanding. Existing approaches can be broadly classified into training-based and training-free methods. Training-based methods

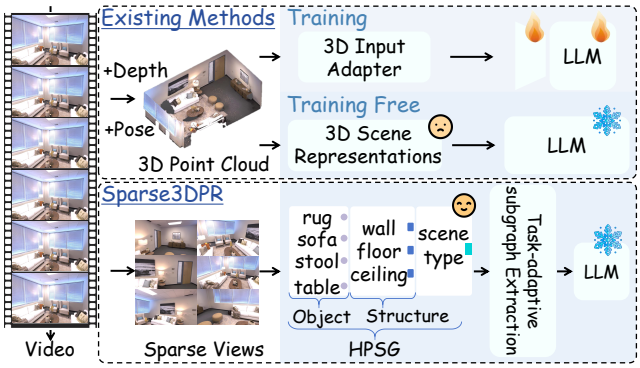

Figure 1: Comparison of 3D scene understanding methods. Unlike existing methods requiring dense 3D inputs or training, Sparse3DPR leverages sparse RGB views to construct a hierarchical plane-enhanced scene graph (HPSG) and performs task-adaptive subgraph extraction for efficient LLM-based scene understanding.

(Fu et al. 2024; Hong et al. 2023; Wang et al. 2023) align 3D geometric or visual features with text features through specialized training, require complex architectures and incur high computational costs. Differently, training-free methods (Gu et al. 2024; Chandhok 2024) build explicit structured representations such as scene graphs (SGs) that encode objects and spatial relations, then convert SGs into textual context for the LLM, thereby leveraging its powerful zero-shot reasoning capability while eliminating training cost.

Despite eliminating 3D-specific training, training-free methods still struggle with reasoning accuracy and computational efficiency in practical deployment. The first bottleneck is the quality and structural organization of 3D scene representations, which serve as the primary contextual input to LLMs and significantly impact reasoning reliability. Current SGs for LLM reasoning fall into two categories: flat SGs, such as ConceptGraphs (Gu et al. 2024), which encode all objects and pairwise relations in a single layer, lacking a hierarchy to organize their massive number of nodes, leading to redundant and token-inefficient inputs. Hierarchical SGs, like TB-HSU (Xu et al. 2025), group objects by their functionality. While offering more structure, their hierarchy disrupts the scene's natural spatial proximity, breaking physi-

*These authors contributed equally.
†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

cal coherence and introducing reasoning ambiguity. Another critical issue is the contextualization method for reasoning. The approach in prior work, such as SceneGPT (Chandhok 2024), adopts a static, non-adaptive approach that converts the entire unfiltered scene representation into a single context for each query. This approach increases computational cost and degrades reasoning accuracy, as task-relevant details are easily obscured by irrelevant information.

To address the challenges, we propose Sparse3DPR, a novel training-free framework for 3D scene understanding from sparse-view RGB inputs, as illustrated in Figure 1. Inspired by human cognitive mechanisms for organizing scenes (Epstein and Kanwisher 1998), Sparse3DPR constructs a hierarchical plane-enhanced scene graph (HPSG) with a geometrically grounded hierarchy anchored by dominant planes (e.g., walls, floors, ceilings), achieving structural efficiency and spatial coherence to overcome the limitations of prior SGs. Semantically, it leverages pretrained vision-language models (VLMs) to enrich fine-grained objects with rich concepts, thereby enabling open-vocabulary capabilities. Furthermore, to effectively leverage this rich and effective representation for LLM-based reasoning, we introduce a task-adaptive subgraph extraction method. It mimics human selective attention by dynamically retrieving only the query-relevant parts of the HPSG. This reduces contextual noise and enables more accurate and efficient task-specific reasoning. To validate the effectiveness of Sparse3DPR, we evaluate it on 3D question answering (QA), a representative task that assesses its unified ability to comprehend language, perceive semantics, and reason about complex spatial relationships within 3D scenes. We further report the average inference time per query as a metric of reasoning efficiency.

The primary contributions of Sparse3DPR are succinctly summarized as follows:

- We design a hierarchical plane-enhanced scene graph that incorporates dominant planar structures as the spatial anchors, which provides a more human-intuitive representation, achieves 6.7%/13.8% EM@1 improvement and 13.0%/7.8% faster than flat scene graphs (Gu et al. 2024) and affordance-based hierarchical scene graphs (Xu et al. 2025) on the 3D QA task.
- We propose a novel task-adaptive subgraph extraction method that dynamically prunes the full scene graph to generate a query-relevant subgraph, which provides the LLM with a focused, noise-free context, thereby enhancing 3D scene reasoning speed and accuracy.
- By integrating the above innovations, we present a training-free scene understanding framework that requires sparse-view RGB images solely. Our method achieves an improvement of 28.7% EM@1 and a speedup of 78.2% compared to ConceptGraphs, while demonstrating competitive performance on public benchmarks and generalizability in real-world scenes.

Related Work

3D Scene Representations for LLM Reasoning

Recent works using dense 3D representations (e.g., CLIP-based (Radford et al. 2021; Peng et al. 2023; Mohiuddin et al. 2024; Jatavallabhula et al. 2023; Yamazaki et al. 2024)) support open-vocabulary perception but lack structure, hindering compositional reasoning. In contrast, 3D SGs offer a structured, object-centric alternative. Classic methods (Armeni et al. 2019; Wald et al. 2020) build graphs where nodes represent objects or spatial regions and edges encode spatial or semantic relationships. Hydra (Hughes, Chang, and Carlone 2022; Hughes et al. 2024) extends this to real-time hierarchical SGs that incrementally organize environments from high-level structures (e.g., buildings and rooms) to fine-grained entities (e.g., objects and agents). These representations effectively organize spatial and are well-suited for LLM integration, thanks to their structured format and ease of conversion to natural language. For example, TB-HSU (Xu et al. 2025) and ConceptGraphs (Gu et al. 2024) construct SGs that can be integrated with LLMs for reasoning. However, TB-HSU is constrained to a predefined set of semantic categories, which limits its generalization and reasoning flexibility. While ConceptGraphs is capable of building open-vocabulary SGs, its flat graph structure lacks a meaningful hierarchy, limiting efficient and coherent reasoning in LLMs. In contrast, our proposed HPSG incorporates dominant planar structures (walls, floors, ceilings) as spatial anchors, while supporting open-vocabulary capabilities and forming a spatially coherent and semantically rich hierarchy suited for LLM reasoning.

3D Scene Understanding with LLMs

Recent works integrate LLMs into 3D scene understanding, often by training them to align spatial and textual representations on 3D-specific datasets (Hong et al. 2023; Wang et al. 2023; Huang et al. 2024; Chen et al. 2024a; Zhang et al. 2025; Fu et al. 2024; Chen et al. 2024b; Mao et al. 2025). For instance, Scene-LLM (Fu et al. 2024) jointly trains 3D encoders with LLMs, while LL3DA (Chen et al. 2024a) uses fusion modules to align geometric and textual features. Despite their effectiveness, these methods require large-scale 3D datasets and still struggle to fully align geometric structures with natural language representations. Alternatively, training-free approaches (Mohiuddin et al. 2024; Gu et al. 2024; Chandhok 2024) leverage pretrained LLMs and scene representations, enabling reasoning without the need for 3D-specific training. SceneGPT (Chandhok 2024), for instance, employs SGs built from dense RGB-D inputs for LLM-based scene understanding. However, converting the entire SG into a single prompt risks exceeding the LLM's context window, and it also interferes with its reasoning by introducing task-irrelevant information. In contrast, our framework is training-free and relies only on sparse-view RGB images, which eliminates the need for 3D-specific data. Furthermore, its task-adaptive subgraph extraction method provides the LLM with more compact and relevant information by pruning away task-irrelevant entities from the full SG.

Methodology

The proposed Sparse3DPR framework provides a training-free solution for 3D scene understanding from sparse RGB views. It first parses the scene to construct an HPSG as a

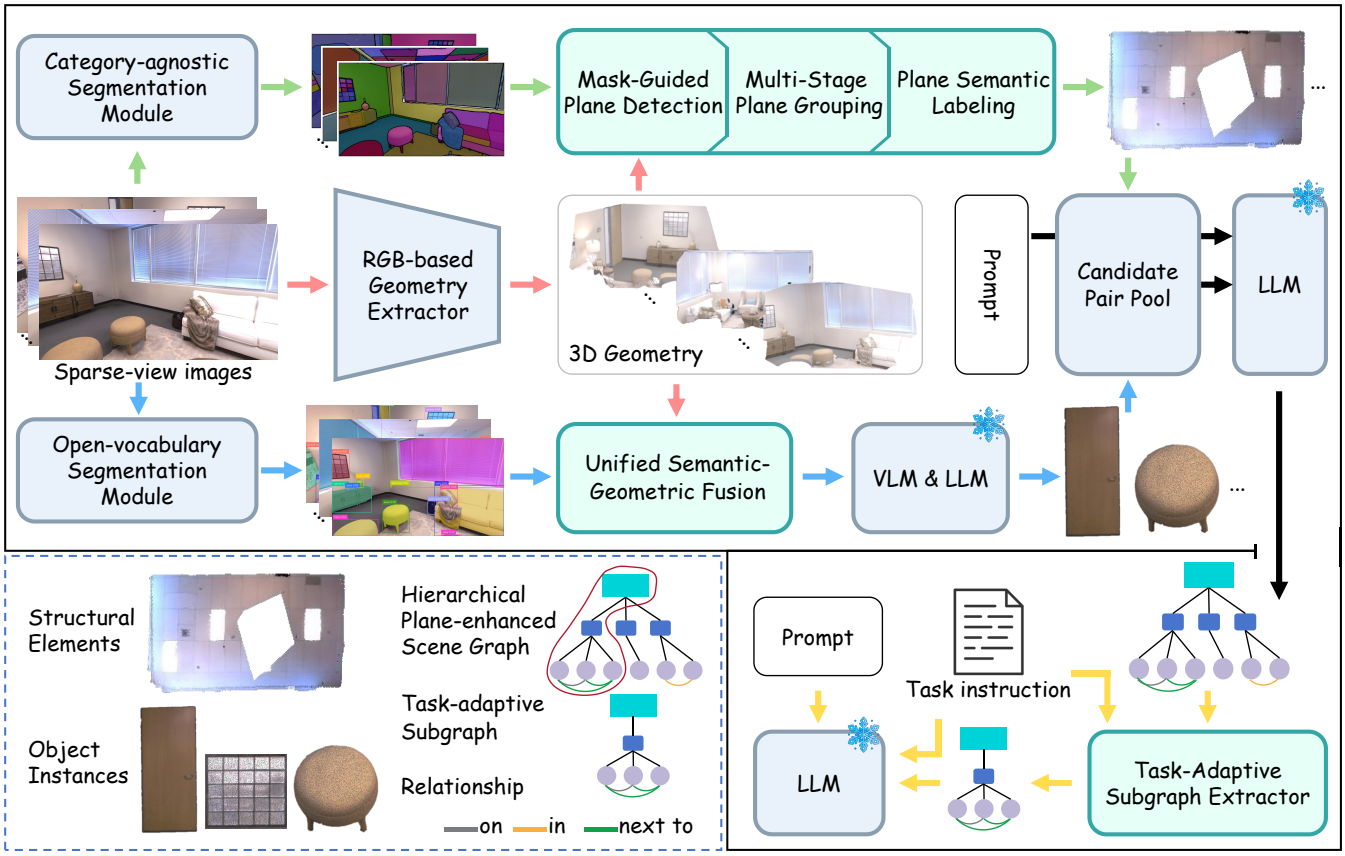

Figure 2: Overview of Sparse3DPR framework. Sparse3DPR is a training-free framework for 3D scene understanding from sparse RGB views. It parses the scene through two branches that integrate 3D geometry: one extracts structural elements by applying class-agnostic masks combined with structural plane detection and labeling, while the other uses open-vocabulary masks for semantic-geometric fusion to obtain object instances and generate captions. These components form a candidate pair pool of topological connections, which are further refined by an LLM to estimate spatial relations between object pairs and construct the HPSG. A task-adaptive subgraph extractor then selects relevant context from the HPSG for LLM-based reasoning.

compact and reasoning-oriented scene representation, and then dynamically prunes the HPSG to focus on query-relevant context via task-adaptive subgraph extraction. The overview of the framework is shown in Figure 2.

Hierarchical Scene Parsing

Geometry Extraction from Sparse RGB Inputs. Effective hierarchical scene parsing requires consistent and geometrically reliable information. Instead of requiring dense, pose-aligned RGB-D inputs, Sparse3DPR utilizes DUSt3R (Wang et al. 2024), a learning-based multi-view reconstruction framework that can infer scene geometry directly from sparse monocular RGB images, even under challenging wide-baseline conditions. To ensure sufficient spatial coverage and viewpoint diversity, we uniformly sample a sparse subset of input images from the available pool, denoted as $\mathcal{I} = \{I_i\}_{i=1}^{n}$. The geometry inference is then formulated as a parameterized function Φ_{D3R}:

$$(\mathcal{X}, \mathcal{C}) = \Phi_{\text{D3R}}(\mathcal{I}; \boldsymbol{\theta}_{\text{D3R}}), \quad P_i = X_i \odot C_i, \tag{1}$$

where $\mathcal{X} = \{X_i\}_{i=1}^{n}$ denotes the set of per-view 3D point maps and $\mathcal{C} = \{C_i\}_{i=1}^{n}$ represents the corresponding confidence maps. $\boldsymbol{\theta}_{\text{D3R}}$ denotes the tunable parameters of Φ_{D3R}. Each $X_i \in \mathbb{R}^{W \times H \times 3}$ encodes the 3D coordinates of every pixel, while each $C_i \in \mathbb{R}^{W \times H}$ provides the associated per-pixel confidence scores. The element-wise product $\odot$ is used to suppress low-confidence points, resulting in a filtered point cloud $P_i \in \mathbb{R}^{N_i \times 3}$ for each view.

Structural Element Extraction. This section details our three-stage pipeline for robustly extracting the foundational structure of the scene (walls, floors, ceilings) from the noisy geometry recovered from sparse views. The pipeline progressively refines an initial set of plane hypotheses into a globally consistent set of structural elements.

The process begins with mask-guided plane detection, where we first leverage SAM2 (Ravi et al. 2024) to generate a set of class-agnostic segmentation masks $\mathcal{M}_i = \{M_i^j\}_{j=1}^{m_i}$ for each input image I_i, with m_i being the number of masks. Each 2D mask M_i^j is lifted into 3D as a candidate point cloud P_i^j by applying it to the inferred point map X_i, weighted by the per-pixel confidence map C_i:

$$P_i^j = X_i \odot (M_i^j \odot C_i). \tag{2}$$

We then by applying a RANSAC-based plane fitting algorithm $\mathcal{R}_{plane}$ to each candidate point cloud P_i^j to generate an initial set of candidate planes:

$$(\mathbf{n}_i^j, d_i^j, \mathcal{O}_i^j) = \mathcal{R}_{plane}(P_i^j; \tau_{\text{dist}}, \rho_{\text{min\_inlier}}), \tag{3}$$

where $\mathbf{n}_i^j$ and d_i^j are the unit normal and offset of the estimated plane, and τ_{dist} is the inlier distance threshold. A candidate plane Π_i^j is considered valid if its inlier ratio $|\mathcal{O}_i^j|/|P_i^j|$ exceeds $\rho_{\text{min\_inlier}}$, where the corresponding inlier set $\mathcal{O}_i^j$ is given by:

$$\mathcal{O}_i^j = \{\mathbf{p} \in P_i^j \mid |\langle \mathbf{n}_i^j, \mathbf{p}\rangle - d_i^j| \leq \tau_{\text{dist}}\}. \tag{4}$$

To merge the initially fragmented and inconsistent candidate planes into a globally coherent set, our pipeline next performs multi-stage plane grouping. This begins with intra-view refinement for each view independently, where we first cluster local candidate planes using DBSCAN in the plane parameter space (PPS), which is defined by the estimated plane parameters $(\mathbf{n}, d)$. This step yields coarse groupings of geometrically co-planar surfaces. To enhance spatial continuity and completeness, these coarse groups are then expanded by a geometry-aware region growing module that incorporates nearby points satisfying the following strict angular and distance constraints:

$$\cos^{-1}(\mathbf{n}_i^{j\top}\mathbf{n}_{\mathbf{p}}) < \theta_{\text{ang}},\ |\mathbf{n}_i^{j\top}\mathbf{p} - d_i^j| < \delta_{\text{dist}}, \tag{5}$$

where $\mathbf{n}_{\mathbf{p}}$ is the pre-computed normal of point $\mathbf{p}$, and θ_{ang} and δ_{dist} are the respective angular and distance thresholds. While this expansion effectively recovers fragmented regions, it can introduce residual over-segmentation. Therefore, a final intra-view DBSCAN is applied to consolidate the expanded planes and merge adjacent, co-planar segments, yielding a refined set of planes Π_i' for each view. These refined per-view sets are then fused in the cross-view alignment stage, where we aggregate all refined sets $\{\Pi_i'\}$ and perform a global DBSCAN clustering in the PPS. This process robustly associates and merges observations of the same plane across different views to produce the globally consistent set of distinct planar surfaces Π_{global}.

The final stage of our pipeline is plane semantic labeling. To distinguish structural elements within Π_{global}, we introduce a geometry-driven semantic labeling method that assigns each plane to a structural category, such as floor, wall, or ceiling, or as a non-structural surface. We begin by detecting floor planes, which are identified by their normal vectors that are closely aligned with the gravity direction, thereby establishing a gravity-aligned coordinate frame for the scene. Using this frame, ceiling planes are classified based on their normals forming an angle of less than 20° with the gravity direction and exhibiting a positive offset d. Wall candidates are identified by normals approximately orthogonal to the floor plane. To distinguish true structural walls from other large vertical surfaces (e.g., cabinet sides), these candidates are then validated through a multi-criteria filtering process. This filter assesses geometric attributes, such as planar area and boundary length, and requires at least two supporting observations from different viewpoints to ensure geometric consistency.

Object Instance Extraction. We introduce an object-centric pipeline that robustly extracts globally consistent 3D object instances from sparse RGB views. It further enriches each instance with descriptive captions and semantic tags.

The process begins with an open-vocabulary instance segmentation module. For each sparse-view image I_i, we first apply the RAM++ model $\mathcal{F}_{\text{T}}$ (Huang et al. 2023) to predict a set of open-vocabulary category labels. These labels are used as text prompts for the GroundingDINO detector $\mathcal{F}_{\text{D}}$ (Liu et al. 2024) to localize candidate object regions. The detected regions are then passed to the SAM2 segmenter $\mathcal{F}_{\text{S}}$ to generate instance masks. To ensure cross-view consistency, we introduce a propagation module $\mathcal{F}_{\text{P}}$ (Cheng et al. 2023), which associates masks of the same object across multiple views. Through this segmentation process, the module generates cross-view consistent instance masks, where each mask is assigned a persistent ID to maintain identity consistency for the object across different viewpoints. The mask generation and ID assignment are jointly formulated as:

$$\left\{(M_i^j, L^j)\right\}_{j=1}^{m_i} = \mathcal{F}_{\text{P}}\left(\mathcal{F}_{\text{S}}\left(\mathcal{F}_{\text{D}}(I_i, \mathcal{F}_{\text{T}}(I_i))\right), \mathbf{S}_{i-1}\right), \tag{6}$$

where m_i is the number of detected instances in view i, M_i^j is the mask for the j-th instance, and L^j is the consistent instance ID shared across views. $\mathbf{S}_{i-1}$ is the propagated state from previous views used to ensure ID consistency.

Building upon the outputs of the previous stage, we perform a unified semantic-geometric fusion. The process begins by combining the 2D object masks from view i with their corresponding 3D geometry from DUSt3R, which yields a set of local object candidates $\hat{o}_i = \{\hat{o}_i^{(1)}, \ldots, \hat{o}_i^{(m)}\}$, where m denotes the number of detected objects. Each candidate $\hat{o}_i^{(m)} = \langle P_i^m, L^m\rangle$ comprises its 3D geometry P_i^m and an associated instance ID L^m. To suppress outliers caused by imprecise mask boundaries, these candidates are subsequently filtered using a density-based clustering method (e.g., DBSCAN), retaining only the dominant cluster as the final geometry. The refined candidates are then progressively merged into a global object set $O = \{o^{(1)}, \ldots, o^{(K)}\}$ based on a unified semantic-geometric association rule. A new detection is merged with an existing object if their IDs match, or if their 3D IoU φ_{geo} exceeds a threshold κ. Otherwise, the detection is registered as a new instance. This fusion logic is formally expressed as:

$$O \leftarrow \begin{cases} \left(O \setminus \{o^{(k)}\}\right) \cup \left\{\langle P^k \cup P_i^m, L^k\rangle\right\} \\ \quad \text{if } \exists\, k,\ \varphi_{\text{geo}} > \kappa \vee L^m = L^k, \\ O \cup \{\langle P_i^m, L^m\rangle\} \quad \text{otherwise.} \end{cases} \tag{7}$$

With the global 3D object set established, the final stage performs vision-language caption generation. We employ a two-stage procedure that first uses a VLM to generate preliminary captions, which are then refined by an LLM. For each object, the process begins by selecting the top-5 views with the highest segmentation confidence. The corresponding object-centered image crops are fed to the VLM with the prompt "*Provide a concise description of the main object in this image.*". This yields a set of preliminary view-specific captions $\widetilde{c}_v = \{\widetilde{c}_{v,1}, \widetilde{c}_{v,2}, \ldots, \widetilde{c}_{v,5}\}$. These preliminary captions are then refined and consolidated by the LLM using

a prompt template $\mathcal{P}$, producing a coherent final caption c_v together with a canonical tag t^* and a candidate tag set $\mathcal{T}$:

$$(c_v, t^*, \mathcal{T}) = \mathrm{LLM}(\widetilde{c}_v, \mathcal{P}). \tag{8}$$

HPSG Construction. Given the set of scene components, including structural plane elements Π and object instances O, we compute a similarity matrix $\mathbf{S}$, where each entry is defined as the 3D bounding box IoU between every pair of components. A minimum spanning tree (MST) is then estimated from $\mathbf{S}$ to establish a candidate pool of topological connections among all components. We further estimate the spatial relations such as 'on', 'in', and 'next_to' between object instance pairs found within the candidate pool by feeding their captions and 3D positions to the LLM using a prompt template. In addition, the LLM summarizes captions of all O to infer the global scene type (e.g., office or room), which is then combined with each structural label to generate captions for Π, such as "*This is a* $\{$*wall*$\}$ *in the* $\{$*office*$\}$." We then construct the HPSG, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The node set $\mathcal{V} = \bigcup_{l=0}^{2} \mathcal{V}_l$ comprises three levels of scene components: $\mathcal{V}_0$ corresponds to the global scene type (e.g., office), $\mathcal{V}_1$ to structural plane elements, and $\mathcal{V}_2$ to object instances. The edge set $\mathcal{E} = \bigcup_{m=0}^{2} \mathcal{E}_m$ captures the hierarchical and spatial relationships, including scene type-to-structure default links $\mathcal{E}_0$, structure-to-object topological connections $\mathcal{E}_1$, and spatial relations between objects $\mathcal{E}_2$.

Task-Adaptive Subgraph Extraction

To dynamically extract task-relevant subgraphs from the full HPSG, we use each node's caption as a semantic anchor for identifying seed nodes. All node captions $\{e_i\}_{i=1}^{N}$ and the user query q are encoded into their corresponding embeddings using a pre-trained SentenceTransformer $S(\cdot)$ model (Reimers and Gurevych 2019):

$$\mathbf{F}_{e_i} = S(e_i), \quad \mathbf{F}_q = S(q). \tag{9}$$

The semantic alignment between the query and each node is then measured using a query-aware scoring function:

$$S(q, e_i; \tau) = \exp\left(\frac{\mathcal{N}(\mathbf{F}_q) \cdot \mathcal{N}(\mathbf{F}_{e_i})}{\tau}\right), \tag{10}$$

where $\mathcal{N}$ denotes L2 normalization, and τ is a temperature scalar (set to 0.07). To efficiently select the most relevant nodes, we use FAISS (Douze et al. 2024) to retrieve the top-K seed nodes that maximize the alignment score:

$$I^* = \underset{i \in \{1,\ldots,N\}}{\arg\,\mathrm{topK}}(\mathcal{S}(q, e_i; \tau)). \tag{11}$$

To retain sufficient contextual information while preserving the hierarchical structure, the retrieved seed nodes are further expanded with their directly connected neighbors $\mathcal{N}(I^*)$ and second-order neighbors $\mathcal{N}^2(I^*)$, together with their corresponding edges from the original graph. The resulting localized subgraph is defined as:

$$\mathcal{G}_q^* = \mathcal{G}\big[I^* \cup \mathcal{N}(I^*) \cup \mathcal{N}^2(I^*)\big], \tag{12}$$

which serves as a compact task-focused subgraph that preserves essential contextual information.

Experiments

Main Results

We design a multi-perspective evaluation to validate the effectiveness of our framework. We first quantitatively evaluate the accuracy and consistency of the nodes in our HPSG via a 3D semantic understanding experiment, as these nodes serve as the contextual foundation for downstream tasks. We then evaluate the framework's reasoning and understanding capabilities using the 3D QA task, which evaluates its ability to perform spatial and semantic reasoning grounded in complex 3D environments via natural language question answering. Finally, we conduct qualitative experiments in a real-world lab scene to further validate the generalization and practical applicability of Sparse3DPR.

Supervision	Method	F-mIoU (%)	mAcc (%)
Training	CLIP (rd64-uni)	39.84	28.21
	CLIP (rd64-uni-refined)	13.00	13.00
	LSeg	51.54	33.39
	OpenSeg	53.74	41.19
Training Free	MaskCLIP	0.94	4.53
	Mask2Former + Global CLIP feat	13.11	10.42
	ConceptFusion	31.31	24.16
	ConceptFusion + SAM	38.70	31.53
	HOV-SG + ViT-H-14	38.60	30.40
	ConceptGraphs^{R}	34.68	37.52
	ConceptGraphs-Detector^{R}	34.70	**37.97**
	Sparse3DPR (Ours)	**39.71**	35.12

Table 1: 3D semantic understanding on Replica. Results marked with ^{R} indicate those reproduced by our implementation. All other results are quoted from their respective publications. Training-based methods involve task-specific fine-tuning, while training-free methods leverage pre-trained models without this requirement.

3D Semantic Understanding. We evaluate the 3D semantic understanding on the Replica (Straub et al. 2019), which contains five office scenes and three apartment rooms. As shown in Table 1, Sparse3DPR is compared with several training-based methods (Lüddecke and Ecker 2022; Li et al. 2022; Ghiasi et al. 2022) and training-free methods (Zhou, Loy, and Dai 2022; Jatavallabhula et al. 2023; Werby et al. 2024; Gu et al. 2024). Sparse3DPR achieves the highest F-mIoU among all training-free baselines and even outperforms several training-based approaches, while also delivering competitive mAcc scores. This performance stems from the semantic-geometric fusion employed during hierarchical scene parsing. Sparse3DPR first establishes a robust geometric foundation by applying DBSCAN to suppress noise in the initial segments, then progressively aligns cross-view features under both geometric and semantic constraints. This process yields a globally consistent scene representation which is essential for supporting precise and reliable reasoning in downstream 3D scene understanding tasks.

3D Scene Question Answering. In this section, we evaluate Sparse3DPR on the ScanQA validation set (Azuma et al. 2022), compare it with state-of-the-art (SOTA) approaches (Azuma et al. 2022; Li et al. 2024; Hong et al. 2023; Chen

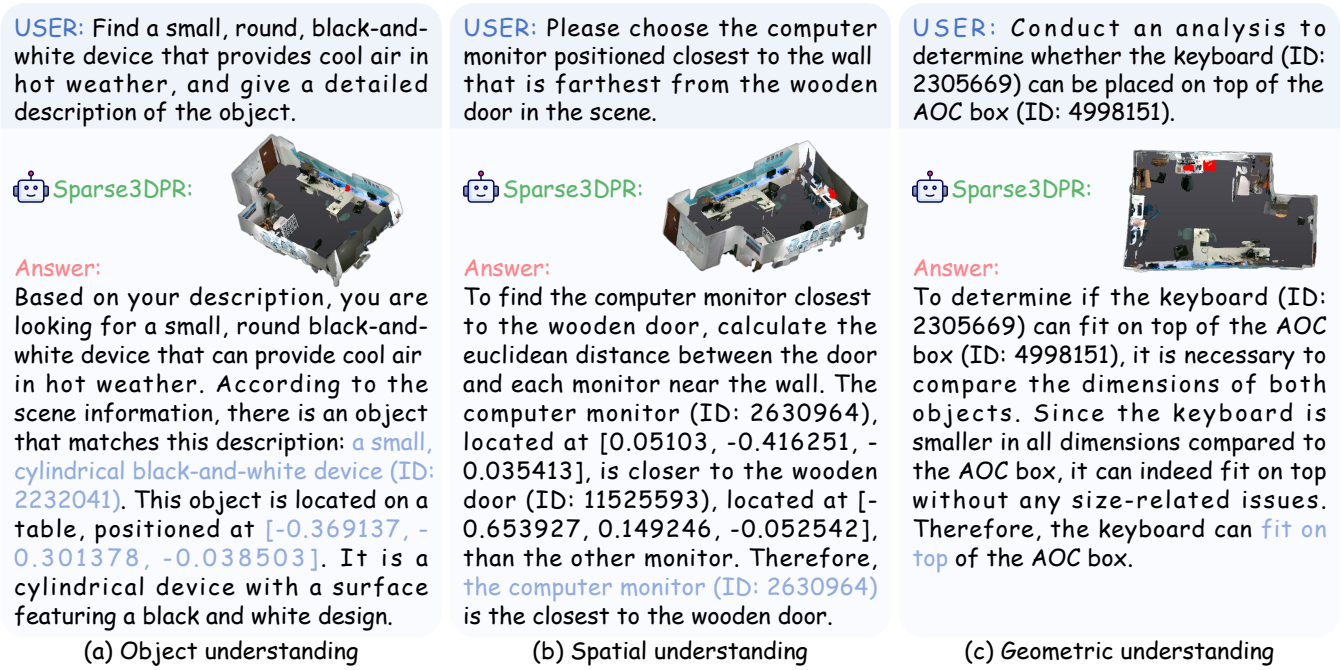

Figure 3: Qualitative results. We showcase Sparse3DPR performing object, spatial, and geometric understanding in a lab scene and locating target objects within the scene. These examples demonstrate its ability to adapt to diverse tasks and generalize to complex real-world indoor scenes.

et al. 2024a; Fu et al. 2024; Huang et al. 2024; Jin et al. 2023; Chen et al. 2024b; Wang et al. 2023) that rely on 3D-specific inputs like point clouds. As shown in Table 2, Sparse3DPR demonstrates highly competitive performance. It achieves SOTA results on several key metrics, including an EM@1 of 27.22% and a CIDEr score of 88.07%, which respectively reflect its factual accuracy and ability to generate semantically rich responses. Notably, this strong performance is achieved in a training-free, zero-shot setting using only sparse RGB inputs. This success is attributed to our framework's ability to provide the LLM with a superior context, where our HPSG establishes a robust and spatially coherent structural foundation, and our task-adaptive subgraph extraction ensures this context is query-relevant and redundancy-free for more accurate and reliable reasoning.

Qualitative Analysis of 3D Scene Understanding. To demonstrate the practical applicability and generalization of our method, we conduct qualitative experiments in a real-world laboratory scene characterized by cluttered object distributions and irregular layouts. This environment presents significant challenges for prior paradigms that are often constrained by domain gaps or a reliance on 3D-specific data such as point clouds or dense posed RGB-D sequences. As demonstrated in Figure 3, Sparse3DPR effectively handles complex queries spanning object understanding, spatial reasoning, and geometric analysis under these conditions. This robust performance stems from the framework's ability to efficiently parse the scene, extract task-relevant context such as captions, 3D positions, and geometric properties, and integrate this information into structured prompts for context-aware reasoning. Crucially, the entire process operates without task-specific fine-tuning and relies solely on sparse-view RGB inputs, demonstrating strong generalization and practical deployability. For more extensive qualitative examples, please refer to the supplementary material.

Ablation Study

To evaluate the impact and effectiveness of HPSG and our task-adaptive subgraph extraction, we conduct ablation studies on 3D QA using a seven-scene subset of Space3D-Bench. For this analysis, we build flat SG baselines using the SG from ConceptGraphs, and further reorganize HPSG into two alternative designs for comparison: a flat structure consistent with ConceptGraphs and an affordance-based hierarchical structure following TB-HSU (Xu et al. 2025). To ensure a fair comparison, all variants are evaluated using the identical reasoning pipeline, including the same LLM and prompt template. Additional details on baselines and experimental settings are available in the supplementary material. The results are summarized in Table 3.

Impact of SG type. To isolate the impact of different SGs on reasoning performance, we disable the task-adaptive subgraph extraction and feed the full graph to the LLM. As shown in the top block of Table 3, Sparse3DPR* (Flat SG) outperforms the baseline ConceptGraphs* (Flat SG) with higher EM@1, indicating the high quality of our underlying scene parsing. In contrast, Sparse3DPR† (Afford. SG) employs a function-centric hierarchy that shifts the LLM's focus from explicit objects and spatial relations to object affordances, improving fluency and reasoning efficiency at the

Supervision	Method	Input	EM@1 ↑	B-1 ↑	B-2 ↑	B-3 ↑	B-4 ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑
Training	ScanQA	PC	21.05	30.24	20.40	15.11	10.08	13.14	33.33	64.86
	3DMIT		13.04	27.63	-	-	5.24	10.70	26.22	48.03
	3D-LLM(Flamingo)		20.40	30.30	17.80	12.00	7.20	12.20	32.30	59.20
	3D-LLM(BLIP2-flant5)		20.50	39.30	25.20	18.40	12.00	14.50	35.70	69.40
	3D-VisTA		22.40	-	-	-	10.40	13.90	35.70	69.60
	LL3DA		-	-	-	-	13.53	15.88	37.31	76.79
	Scene-LLM		27.20	**43.60**	26.80	19.10	12.00	16.60	40.00	80.00
	Chat-Scene		21.62	43.20	**29.06**	20.57	14.31	**18.00**	**41.56**	87.70
	3D-VLP		21.65	30.53	21.33	16.67	11.15	13.53	34.51	66.97
	Grounded 3D-LLM		-	-	-	-	13.40	-	-	72.70
	Chat-3D		-	29.10	-	-	6.40	11.90	28.50	53.20
	Chat-3D v2		21.10	38.40	-	-	7.30	16.10	40.10	77.10
Training Free	**Sparse3DPR(Ours)**	SVI	**27.22**	36.23	28.42	**22.92**	**14.99**	17.40	37.98	**88.07**

Table 2: Evaluations on ScanQA. Our training-free method, operating on sparse-view images (SVI), is benchmarked against training-based approaches that use point clouds (PC). Best results are highlighted in bold. B-1 to B-4 are BLEU-n scores. Sparse3DPR achieves state-of-the-art results on several metrics (EM@1, BLEU-3, BLEU-4, and CIDEr).

Method	Subgraph	SG Type	EM@1 ↑	B-1 ↑	B-2 ↑	B-3 ↑	B-4 ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑	Inference Time (avg) ↓
ConceptGraphs	✗	Flat	20.23	11.72	6.64	3.79	2.18	11.05	35.54	85.43	2.15s
ConceptGraphs*	✗	Flat	26.94	20.92	12.11	7.38	4.47	15.82	49.69	115.79	1.47s
Sparse3DPR*	✗	Flat	28.97	21.42	12.22	7.55	4.55	15.96	49.90	133.38	1.08s
Sparse3DPR†	✗	Afford.	27.17	21.51	12.69	8.02	4.99	15.45	47.67	122.30	1.02s
Sparse3DPR*	✗	HPSG	**30.91**	**24.03**	**13.98**	**8.86**	**5.64**	**16.51**	**51.81**	**141.15**	**0.94s**
ConceptGraphs	✓	Flat	28.04	27.71	17.14	10.94	6.18	14.89	49.06	126.00	0.39s
ConceptGraphs*	✓	Flat	32.49	27.14	16.53	10.61	6.46	15.50	52.22	140.49	0.39s
Sparse3DPR*	✓	Flat	32.82	30.43	19.08	12.85	8.16	16.55	52.28	146.39	0.38s
Sparse3DPR†	✓	Afford.	30.78	29.00	17.54	11.41	7.08	15.13	50.98	136.08	0.37s
Sparse3DPR (Ours)	✓	HPSG	**34.68**	**31.13**	**19.90**	**13.60**	**8.87**	**17.56**	**53.17**	**160.73**	**0.32s**

Table 3: Ablation study on Space3D-Bench. All compared methods are training-free. "ConceptGraphs" and "ConceptGraphs*" serve as flat SG baselines constructed following prior methods. Sparse3DPR* denotes variants of Sparse3DPR either without the subgraph extraction method or with flat SGs obtained by flattening HPSG. Sparse3DPR† reorganizes HPSG into an affordance-based hierarchical structure. The *SG Type* column specifies the organization of the scene graph: Flat (non-hierarchical), Afford. (affordance-based hierarchy), or HPSG (hierarchical plane-enhanced scene graph).

cost of accuracy (EM@1). Our Sparse3DPR* (HPSG), however, uses a physically-grounded hierarchy that preserves spatial coherence, leading to significant improvements in both EM@1 accuracy and reasoning speed. This proves that a human-intuitive, spatially coherent SG structure is more effective for LLM-based reasoning.

Impact of task-adaptive subgraph extraction. We first analyze the primary impact of our task-adaptive subgraph extraction. As shown in Table 3 (bottom vs. top blocks), enabling this method yields a significant and consistent performance improvement across all SG types in both reasoning accuracy and efficiency. These results demonstrate that dynamically providing the LLM with a focused, query-relevant, and noise-free context is critical for effective reasoning. Furthermore, these results also reveal an interplay between the subgraph extraction method and the underlying SG structure (as shown in the bottom block). The performance benefit is most pronounced when combined with our HPSG, which achieves the best results across all metrics (e.g., 0.32s reasoning time). In contrast, when the subgraph is combined with the Affordance-based SG, performance declines. We attribute this to its function-centric hierarchy, which emphasizes functional relationships, leading to inaccurate seed node selection and the inclusion of irrelevant scene context. By comparison, HPSG preserves spatial coherence through dominant planar structures as spatial anchors, enabling precise seed node retrieval and producing cleaner, task-relevant subgraphs for improved performance.

Conclusion

We propose Sparse3DPR, a framework that addresses accuracy and efficiency challenges in the practical deployment of training-free 3D scene understanding. By constructing an HPSG from sparse-view RGB inputs, it provides a spatially coherent, semantically rich, and reasoning-friendly representation, while the task-adaptive subgraph extraction method dynamically filters redundant context and retains task-relevant information, thereby improving reasoning accuracy and efficiency. Experiments show that Sparse3DPR significantly improves both accuracy and speed over previous training-free methods and achieves comparable performance to training-based counterparts on the ScanQA benchmark, confirming its efficiency and generalizability for real-world applications. Future work will focus on extending this framework to temporal reasoning in dynamic environments.

Acknowledgments

This work was supported by the project "Research of Rapid 3D Digital Acquisition and Reconstruction Technology and Equipment for Cultural Heritage" (No.2024JK4002) and the National Natural Science Foundation of China under Grant Nos. 62572468 and 62402493.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Armeni, I.; He, Z.-Y.; Gwak, J.; Zamir, A. R.; Fischer, M.; Malik, J.; and Savarese, S. 2019. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision (CVPR)*, 5664–5673.

Azuma, D.; Miyanishi, T.; Kurita, S.; and Kawanabe, M. 2022. ScanQA: 3D Question Answering for Spatial Scene Understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19129–19139.

Chandhok, S. 2024. SceneGPT: A Language Model for 3D Scene Understanding. arXiv:2408.06926.

Chen, S.; Chen, X.; Zhang, C.; Li, M.; Yu, G.; Fei, H.; Zhu, H.; Fan, J.; and Chen, T. 2024a. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26428–26438. Seattle, WA, USA.

Chen, Y.; Yang, S.; Huang, H.; Wang, T.; Xu, R.; Lyu, R.; Lin, D.; and Pang, J. 2024b. Grounded 3D-LLM with Referent Tokens. arXiv:2405.10370.

Cheng, H. K.; Oh, S. W.; Price, B.; Schwing, A.; and Lee, J.-Y. 2023. Tracking Anything with Decoupled Video Segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 1316–1326. Paris, France.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The Faiss library. arXiv:2401.08281.

Epstein, R.; and Kanwisher, N. 1998. A Cortical Representation of the Local Visual Environment. *Nature*, 392(6676): 598–601.

Fu, R.; Liu, J.; Chen, X.; Nie, Y.; and Xiong, W. 2024. Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning. arXiv:2403.11401.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, 540–557. Tel Aviv, Israel.

Gu, Q.; Kuwajerwala, A.; Morin, S.; Jatavallabhula, K. M.; Sen, B.; Agarwal, A.; Rivera, C.; Paul, W.; Ellis, K.; Chellappa, R.; et al. 2024. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 5021–5028.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3D-LLM: Injecting the 3D World into Large Language Models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, 20482–20494. New Orleans, LA.

Huang, H.; Chen, Y.; Wang, Z.; Huang, R.; Xu, R.; Wang, T.; Liu, L.; Cheng, X.; Zhao, Y.; Pang, J.; et al. 2024. Chatscene: Bridging 3d scene and large language models with object identifiers. *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*.

Huang, X.; Huang, Y.-J.; Zhang, Y.; Tian, W.; Feng, R.; Zhang, Y.; Xie, Y.; Li, Y.; and Zhang, L. 2023. Open-Set Image Tagging with Multi-Grained Text Supervision. arXiV:2310.15200.

Hughes, N.; Chang, Y.; and Carlone, L. 2022. Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization. In *Robotics: Science and Systems (RSS)*.

Hughes, N.; Chang, Y.; Hu, S.; Talak, R.; Abdulhai, R.; Strader, J.; and Carlone, L. 2024. Foundations of spatial perception for robotics: Hierarchical representations and real-time systems. *The International Journal of Robotics Research*, 43(10): 1457–1505.

Jatavallabhula, K. M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Maalouf, A.; Li, S.; Iyer, G.; Saryazdi, S.; Keetha, N.; Tewari, A.; Tenenbaum, J. B.; de Melo, C. M.; Krishna, M.; Paull, L.; Shkurti, F.; and Torralba, A. 2023. ConceptFusion: Open-set Multimodal 3D Mapping. arXiv:2302.07241.

Jin, Z.; Hayat, M.; Yang, Y.; Guo, Y.; and Lei, Y. 2023. Context-Aware Alignment and Mutual Masking for 3D-Language Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10984–10994. Vancouver, Canada.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven Semantic Segmentation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, Z.; Zhang, C.; Wang, X.; Ren, R.; Xu, Y.; Ma, R.; Liu, X.; and Wei, R. 2024. 3dmit: 3d multi-modal instruction tuning for scene understanding. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 1–5. IEEE.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 38–55. Milan, Italy.

Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-radiology*, 1(2): 100017.

Lüddecke, T.; and Ecker, A. 2022. Image Segmentation Using Text and Image Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7086–7096. New Orleans, LA, USA.

Mao, Y.; Zhong, J.; Fang, C.; Zheng, J.; Tang, R.; Zhu, H.; Tan, P.; and Zhou, Z. 2025. SpatialLM: Training Large Language Models for Structured Indoor Modeling. arXiv:2506.07491.

Mohiuddin, R.; Prakhya, S. M.; Collins, F.; Liu, Z.; and Borrmann, A. 2024. OpenSU3D: Open World 3D Scene Understanding using Foundation Models. arXiv:2407.14279.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. OpenScene: 3D Scene Understanding with Open Vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–824. Vancouver, Canada.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, 8748–8763.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; Clarkson, A.; Yan, M.; Budge, B.; Yan, Y.; Pan, X.; Yon, J.; Zou, Y.; Leon, K.; Carter, N.; Briales, J.; Gillingham, T.; Mueggler, E.; Pesqueira, L.; Savva, M.; Batra, D.; Strasdat, H. M.; Nardi, R. D.; Goesele, M.; Lovegrove, S.; and Newcombe, R. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. arXiv:1906.05797.

Wald, J.; Dhamo, H.; Navab, N.; and Tombari, F. 2020. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3961–3970.

Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2024. DUST3R: Geometric 3D Vision Made Easy. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 20697–20709. Seattle, WA, USA.

Wang, Z.; Huang, H.; Zhao, Y.; Zhang, Z.; and Zhao, Z. 2023. Chat-3D: Data-efficiently Tuning Large Language Model for Universal Dialogue of 3D Scenes. arXiv:2308.08769.

Werby, A.; Huang, C.; Büchner, M.; Valada, A.; and Burgard, W. 2024. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*.

Xu, W.; Ila, V.; Zhou, L.; and Jin, C. T. 2025. TB-HSU: Hierarchical 3D Scene Understanding with Contextual Affordances. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 8960–8968.

Yamazaki, K.; Hanyu, T.; Vo, K.; Pham, T.; Tran, M.; Doretto, G.; Nguyen, A.; and Le, N. 2024. Open-Fusion: Real-Time Open-Vocabulary 3D Mapping and Queryable Scene Representation. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 9411–9417. Yokohama, Japan.

Zhang, J.; Chen, Y.; Zhou, Y.; Xu, Y.; Huang, Z.; Mei, J.; Chen, J.; Yuan, Y.-J.; Cai, X.; Huang, G.; Quan, X.; Xu, H.; and Zhang, L. 2025. From Flatland to Space: Teaching Vision-Language Models to Perceive and Reason in 3D. arXiv:2503.22976.

Zhi, H.; Chen, P.; Li, J.; Ma, S.; Sun, X.; Xiang, T.; Lei, Y.; Tan, M.; and Gan, C. 2025. Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 3761–3771.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 696–712. Tel Aviv, Israel.