RAG-3DSG: Enhancing 3D Scene Graphs with Re-Shot Guided Retrieval-Augmented Generation

Yue Chang[1], Rufeng Chen[1], Zhaofan Zhang[1],
Yi Chen[2], Yifan Tian[1], and Sihong Xie[1⋆]

[1] The Hong Kong University of Science and Technology (Guangzhou), China
[2] Jilin University, China
ychang500@connect.hkust-gz.edu.cn, sihongxie@hkust-gz.edu.cn

Abstract. Open-vocabulary 3D Scene Graph (3DSG) can enhance various downstream tasks in robotics by leveraging structured semantic representations, yet current 3DSG construction methods suffer from semantic inconsistencies caused by noisy cross-image aggregation under occlusions and constrained viewpoints. To mitigate the impact of such inconsistency, we propose **RAG-3DSG**, which introduces re-shot guided uncertainty estimation. By measuring the semantic consistency between original limited viewpoints and re-shot optimal viewpoints, this method quantifies the underlying semantic ambiguity of each graph object. Based on this quantification, we devise an Object-level Retrieval-Augmented Generation (RAG) that leverages low-uncertainty objects as semantic anchors to retrieve more reliable contextual knowledge, enabling a Vision-Language Model to rectify the predictions of uncertain objects and optimize the final 3DSG. Extensive evaluations across three challenging benchmarks and real-world robot trials demonstrate that RAG-3DSG achieves superior recall and precision, effectively mitigating semantic noise to provide more reliable scene representations for robotics tasks.

Keywords: 3D Scene Graph · Uncertainty Estimation · RAG

1 INTRODUCTION

Compact and expressive representation of complex and semantic-rich 3D scenes has long been a fundamental challenge in robotics, with direct impact on downstream tasks such as robot manipulation [9, 28, 31, 40] and navigation [3, 5, 30, 36, 40, 42]. A promising representation is 3D scene graphs (3DSGs) [2, 6], which encode a scene into a graph where nodes denote objects and edges capture their pairwise relationships. Early efforts focus on building 3DSGs by detecting objects and relationships from a closed vocabulary [10,29,37]. While these methods perform well and are efficient in fixed environments, their reliance on a closed vocabulary restricts their generalization to novel, unseen scenes in the complex real-world. To mitigate this limitation, recent approaches [8,12,18,21,25,36,39,40,43] leverage vision-language foundation models to provide open-vocabulary 3DSG generation and produce more expressive representations for diverse scenes.

⋆ Corresponding author.

Despite this progress, open-vocabulary approaches largely adhere to the one-way pipeline of per-image object-level information extraction followed by cross-image aggregation. However, as illustrated in the upper part of Fig. 1, constrained viewpoints, occlusions, and other poor imaging conditions can introduce a significant amount of noise, leading to semantic inconsistency where the same object is described inconsistently across different views. For example, as shown in Fig. 1 (b), although the object of interest is the *table*, the presence of an occluding *vase* leads to multiple crop captions being mistakenly recognized as *vase*. Crucially, existing methods often blindly aggregate these conflicting captions or embeddings across images without discrimination, allowing high-noise observations to corrupt the accuracy of the resulting 3DSG. Such noise-induced inaccuracies in 3DSGs are unacceptable for downstream robotic tasks, particularly in safety-critical scenarios. For example, if a medication bottle is incorrectly

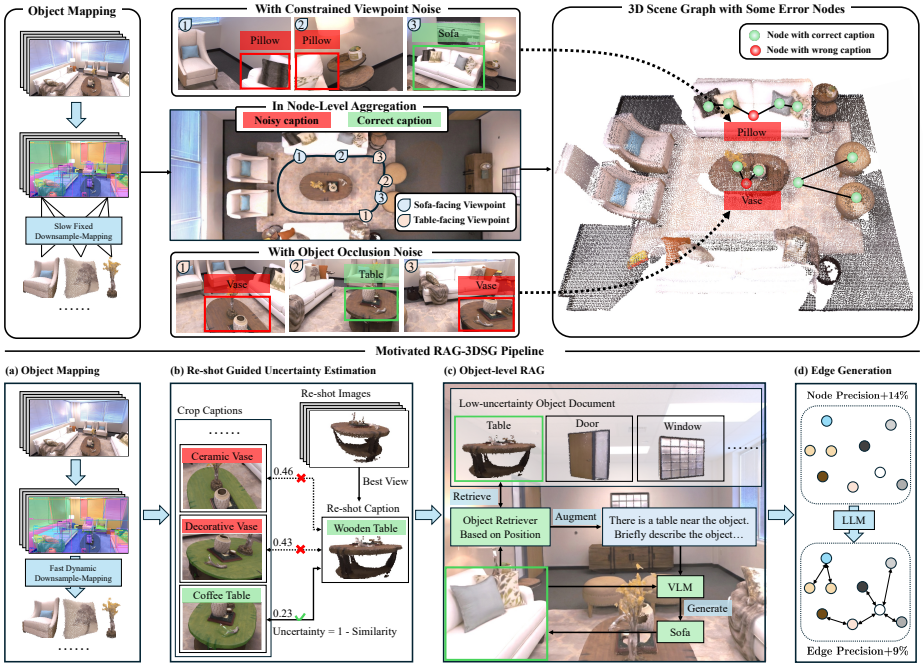

Fig. 1: An overview of the RAG-3DSG framework. The upper part illustrates common challenges in multi-view 3D scene graph generation. Our pipeline addresses these issues through **(a)** Multi-view RGB-D frames are segmented and fused into a global object list with point clouds and semantic embeddings. **(b)** Re-shot images are used to select the best-view caption, which is compared with crop captions to estimate uncertainty; clustering is applied and the top-1 cluster is retained for fusion. **(c)** Low-uncertainty objects form a retrieval document, while high-uncertainty objects leverage retrieved context for caption refinement via a VLM. **(d)** Finally, spatial and semantic relationships among objects are inferred by an LLM to construct the 3D scene graph.

described as a beverage container, a service robot could deliver dangerous substances to humans. Therefore, rather than blindly trusting aggregated features, we argue that the system must "know what it doesn't know"—quantifying semantic ambiguity to distinguish reliable predictions from inconsistent ones.

To this end, we propose **RAG-3DSG**, a framework that mitigates aggregation noise through a two-phase "diagnose-then-rectify" paradigm. The core challenge lies in breaking the dependency on finite input trajectories; while existing methods [7, 15, 21] filter for the best available object frames, they remain "passive observers" restricted to captured viewpoints—even when the true clear object was never recorded. To transform this limited observation into active verification [34, 44, 45], our first phase conceptually shifts the system's role to an *active photographer*. Since reconstructed 3D geometry naturally disentangles objects from background clutter, we virtually 're-shoot' objects from optimal angles. By comparing captions generated from these unconstrained re-shots against the original passive crops (Fig. 1 (b)), we quantify the semantic ambiguity of each object. However, identifying uncertainty is only half the battle; rectifying it without hallucination requires robust context. Therefore, in the second phase, we use these uncertainty scores to drive our Object-level Retrieval-Augmented Generation (RAG) module. Instead of forcing the Vision-Language Model (VLM) to guess semantics in isolation, low-uncertainty objects are anchored as reliable context. High-uncertainty objects then trigger the retrieval of surrounding low-uncertainty "semantic anchors" on the 3DSG to guide object caption refinement (Fig. 1 (c)), which iteratively improves the overall 3DSG (Fig. 1 (d)).

To validate the effectiveness of our approach, we conduct extensive experiments on two challenging real-world benchmarks, SceneFun3D [4] and FunGraph3D [43], alongside a rigorous fine-grained human evaluation on the Replica [32] dataset. Quantitative results demonstrate that RAG-3DSG establishes a new state-of-the-art in both object classification and relationship prediction. Furthermore, human evaluation confirms that our method yields significantly cleaner and more precise scene graphs, substantially suppressing duplicate node predictions and relational hallucinations. Finally, robust performance under injected tracking noise on real-world datasets, coupled with successful real-world robotic experiments, demonstrates that our active re-shot mechanism effectively withstands noise, underscoring its readiness for reliable real-world deployment.

2 RELATED WORK

Scene Graph Generation (SGG) Scene graphs were initially introduced in the 2D domain to extract objects and their relationships from images, providing a structured representation that supports a highly abstract understanding of scenes for intelligent agents [14, 19, 23, 24, 33, 41]. A scene graph is typically composed of three fundamental elements—objects, attributes, and relations—and can be expressed as a set of visual relationship triplets in the form of <subject, relation, object> [20]. In the field of 3D scene representations, several works [2, 16, 35] have drawn inspiration from 2D scene graphs and extended the

concept into the 3D domain. Compared to dense, hard-to-interpret point clouds with per-point semantics, 3D scene graphs (3DSGs) provide a more compact and structured abstraction of the scene. By organizing objects and their relationships into a graph representation, they enable more efficient reasoning and facilitate downstream tasks such as robotic navigation [3,5,30,36,40,42] and scene understanding [1,27]. Early efforts in 3D scene graph generation (3DSGG) enabled the construction of real-time systems capable of dynamically building hierarchical 3D scene representations [10, 29, 37]. However, these methods were strictly confined to the closed-vocabulary setting, which restricted their applicability to a limited range of tasks. More recently, several works [8, 12, 18, 21, 25, 36, 39, 40, 43] have begun to explore open-vocabulary approaches for 3D scene graph generation. Using Vision Language Models (VLMs) and Large Language Models (LLMs), these methods sacrifice part of the real-time capability but significantly expand the range of object categories and relations that can be recognized, thus broadening the applicability to a wider variety of complex downstream tasks.

Open-vocabulary 3DSGG Recent advances have extended 3DSGG to the open-vocabulary setting by leveraging VLMs and LLMs. Existing approaches can be divided into two main paradigms. The first paradigm follows a caption-first strategy, where objects in each image are independently described and later aggregated into scene-level semantics. In practice, multiple masked views of the same object are captioned separately, and a LLM is then used to aggregate these descriptions into a final caption (e.g., [8]). The second paradigm [12,18,36, 39] adopts an embedding-first strategy, where semantic embeddings of multiple masked views of the same object are first extracted and aggregated without explicit captioning. The aggregated embeddings are then converted into captions or directly aligned with user queries in a joint vision–language space using models such as CLIP [26] or SEEM [46]. In addition, some works assign captions to object embeddings by matching them against an arbitrary vocabulary with CLIP, a strategy commonly referred to as open-set 3DSG [25]. Despite their differences, both paradigms share the same underlying pipeline of per-image information extraction followed by cross-image information aggregation. However, the information extraction process is often affected by constrained viewpoints and object occlusion, which introduce noise into the representations. As a result, aggregated information may be inaccurate, leading to reduced precision in the node and edge captions of the 3DSGs. This limitation cannot be simply resolved by adopting more powerful captioning models, as the noise originates from inherent challenges in multi-view perception such as occlusion and viewpoint constraints.

Viewpoint-Aware 3D Semantic Reasoning The quality of semantic understanding is intrinsically linked to observation viewpoints [34]. To mitigate occlusion in complex scenes, recent 3DSG methods [7,15,21] have adopted frame-selection strategies to ensure robust object captioning. For instance, BBQ [21] clusters camera poses to select the frame with the maximum projected area for description generation, while MomaGraph [15] filters the input stream for significant keyframes to obtain informative visual context. Despite their effectiveness, these approaches remain passive: they are strictly bound by the finite input

trajectory, selecting only the "best of the available" views even if the optimal angle was never captured. In contrast, emerging research in the broader field of general 3D perception advocates active spatial reasoning using 3D geometry. Think3D [44] proposes an interactive framework where agents actively manipulate viewpoints, transforming spatial reasoning into a "3D Chain-of-Thought" process. Similarly, Chain-of-View (CoV) [45] converts VLMs into active viewpoint reasoners, employing a coarse-to-fine exploration strategy to expand "finite observations" into an "infinite" verification space. Our RAG-3DSG bridges this gap by introducing active verification into open-vocabulary 3DSG generation.

3 METHOD

In this section, we present RAG-3DSG, a framework designed to bridge the gap between passive perception and active verification in open-vocabulary 3D scene graph generation. As illustrated in Fig. 1, our pipeline consists of three main stages: (1) **Cross-image Object Mapping** (Sec. 3.1), where we incrementally fuse local observations into a global object list, employing dynamic downsampling to balance representation efficiency with point cloud density; (2) **Node Caption Generation** (Sec. 3.2), which implements our diagnose-then-rectify mechanism: utilizing re-shot guided uncertainty estimation to identify semantic ambiguity, followed by Object-level RAG to rectify unreliable predictions; (3) **Edge Caption Generation** (Sec. 3.3), where we leverage the refined object nodes to prompt an LLM for interpretable inter-object relationship generation.

3.1 Cross-image Object Mapping

Given an RGB-D image sequence $I = \{I_1, \ldots, I_t\}$ with poses $P = \{P_1, \ldots, P_t\}$, we maintain a global object list O^{global} by incrementally fusing local objects. This process comprises local information extraction followed by global fusion.

Local Information Extraction For each incoming frame I_t, we first extract local object instances. A class-agnostic segmentation model SAM$(\cdot)$ [17] generates 2D masks $\{m_{t,i}\}$, which are then encoded by CLIP$(\cdot)$ [26] to obtain semantic embeddings $\{f_{t,i}\}$. Using the camera pose P_t, we project these masks into 3D space to form point clouds $\{p_{t,i}\}$. This yields a local object list $O_t^{\text{local}} = \{\langle f_{t,i}^{\text{local}}, p_{t,i}^{\text{local}} \rangle\}_{i=1}^{M}$, serving as the raw input for global fusion.

Global Fusion with Dynamic Downsampling To integrate O_t^{local} into the global list O_{t-1}^{global}, we employ an incremental fusion pipeline. However, standard fusion with fixed-size voxelization often leads to a trade-off: it either over-samples large background structures or under-samples small objects (losing details crucial for our subsequent re-shot verification). To address this, we apply a scale-adaptive downsampling strategy prior to fusion. The voxel size $\delta_{t,i}^{\text{voxel}}$ for the i-th

object in the t-th frame is dynamically adjusted based on its spatial extent:

$$\delta_{t,i}^{\text{voxel}} = \delta^{\text{sample}} \cdot \|\text{Bbox}(p_{t,i})\|_2^{1/2}, \tag{1}$$

where δ^{sample} is a base constant and $\|\text{Bbox}(p_{t,i})\|_2$ is the diagonal length of the object's 3D bounding box. This strategy ensures that small objects retain high point density for reliable re-identification, while large objects remain efficient.

Following dynamic downsampling, we associate local object $o_{t,i}^{\text{local}}$ with global object $o_{t-1,j}^{\text{global}}$ by maximizing a composite similarity score $\theta = \theta_{\text{sem}} + \theta_{\text{spa}}$. The semantic similarity θ_{sem} is the cosine similarity between the CLIP embeddings. Spatial similarity θ_{spa} is defined as the *dynamic nearest neighbor ratio* to measure point cloud overlap. Rather than a fixed radius, its matching threshold scales dynamically with object sizes (derived from Eq. (1)). Pairs exceeding a similarity threshold δ^{sim} are merged: the point clouds are unified ($p^{\text{global}} \leftarrow p^{\text{global}} \cup p^{\text{local}}$), and embeddings are updated via a moving average to accumulate semantic information. Unmatched local objects are registered as new global instances.

3.2 Node Caption Generation

Having constructed the global object list, we now enter the core active verification phase. Instead of passively accepting the noisy semantics from input frames, we implement a diagnose-then-rectify mechanism. As shown in Fig. 1, we first derive initial captions from passive observations, and then introduce a re-shot guided uncertainty estimation module. By actively synthesizing "best-view" re-shots to verify the initial captions, we rigorously quantify semantic ambiguity, separating reliable anchors from uncertain nodes. These verified anchors serve as the knowledge base for our Object-level RAG, which retrieves spatial context to rectify the ambiguous predictions, ensuring globally consistent node semantics.

Initial Caption Generation: The Passive Baseline For each global object $o_{t,i}^{\text{global}}$, we first extract semantic information from the original input sequence. We maintain the top-k views with the highest segmentation confidence and feed these object-level crops into a VLM [11] to obtain initial captions $c_{t,i} = \{c_1, \ldots, c_k\}$. However, as illustrated in Fig. 1 (Top), this process is inherently passive: it is strictly bound by the camera's trajectory. If the camera only captures an object from a constrained or occluded viewpoint, the resulting captions $c_{t,i}$ will inevitably be noisy or incorrect, necessitating an active verification mechanism.

Re-shot Guided Uncertainty Estimation: Active Verification To break the dependency on passive input views, we propose an active re-shot strategy. Conceptually, we transform the role of the observer from a passive recorder to an active photographer. Since the reconstructed point cloud p aggregates temporal observations into a holistic geometry, it is disentangled from transient occlusion and background clutter, offering an unconstrained verification space. Our goal

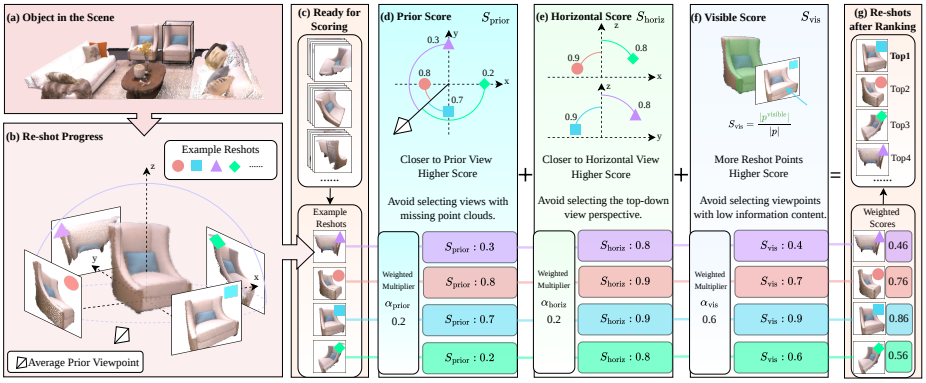

Fig. 2: Workflow of the Re-shot Guided Uncertainty Estimation. **(a-b)** We reconstruct the object and simulate active re-shots on a sampling hemisphere. **(c-g)** A robust multi-factor scoring system ($S_{\text{prior}}, S_{\text{horiz}}, S_{\text{vis}}$) evaluates each candidate view to identify the most informative angle, effectively avoiding ambiguity caused by extreme perspectives.

is to simulate an active optimal photographer that "re-shoots" the object from a scientifically determined "best view" to capture its underlying semantics.

As depicted in Fig. 2 (b), we uniformly sample 64 virtual candidate cameras on a hemisphere centered at the target object o. To identify the most informative view among these candidates, we design a robust view quality scoring system that mimics human preference for clear visual observation. As shown in Fig. 2 (d-f), the score for a candidate view c_i comprises three complementary terms:

$$S_{\text{vis}} = \frac{|p^{\text{visible}}|}{|p|}, \quad S_{\text{horiz}} = 1 - |v_i \cdot g|, \quad S_{\text{prior}} = \tfrac{1}{2}(1 + v_i \cdot f). \tag{2}$$

Specifically, S_{vis} maximizes information content by preferring views with high point visibility; S_{horiz} avoids disorientation by penalizing top-down or bottom-up views that are aligned with the gravity vector g (i.e., the downward z-axis in the gravity-aligned world frame); and S_{prior} maintains consistency with the average prior viewpoint f (the average viewing direction of the source frames observing the object) to avoid complete hallucinations from unseen backsides. The weighted sum $S_{\text{view}} = \alpha_{\text{vis}} S_{\text{vis}} + \alpha_{\text{horiz}} S_{\text{horiz}} + \alpha_{\text{prior}} S_{\text{prior}}$ (subject to $\sum \alpha = 1$) guides us to render the optimal re-shot image I^{reshot} (see Fig. 2 (g)). Importantly, since candidate scoring relies purely on lightweight operations and only the single top-scoring view triggers rendering and VLM querying, this active verification introduces minimal computational overhead.

With the optimal re-shot caption c^{reshot} generated from I^{reshot}, we quantify the object reliability by comparing this "active verification" against the "passive observations". We compute the cosine similarities between their embeddings:

$$\{s_1, \ldots, s_k\} = \left\{\cos\left(\text{CLIP}(c^{\text{reshot}}), \text{CLIP}(c_i)\right)\right\}_{i=1}^{k}. \tag{3}$$

High similarity implies that the initial captions align with the reshot one, indicating high reliability. Conversely, low similarity signals conflict, flagging the object as uncertain. Specifically, these similarity scores are clustered using the KMeans algorithm with $K = 3$. The cluster exhibiting the highest mean similarity is designated as the consensus cluster. The initial passive captions within this consensus cluster, together with the active re-shot caption c^{reshot}, are then aggregated by an LLM to generate the final consensus caption $\hat{c}$. Accordingly, the average similarity score of this consensus cluster is derived as the object's reliability score $\hat{s}$, which serves as the criterion for the subsequent RAG.

Object-level RAG: Contextual Rectification Following the uncertainty diagnosis, we proceed to the rectification phase. We first rank all objects by their uncertainty scores $1 - \hat{s}$ and apply a prompt to the VLM [11] to filter out background objects via crops. The top-50% low-uncertainty objects are included in the object document for RAG, where each final caption c is set to $\hat{c}$. For the remaining high-uncertainty objects, we perform refinement using contextual information. Specifically, a 3D position-based retriever retrieves the nearest object in the document, whose caption c^{env} serves as augmented auxiliary context. In addition, we construct a composite image by concatenating the re-shot image (providing global context) with the crop image that yields the highest similarity score. This composite image, together with a text prompt containing c^{env}, is fed into a VLM [11] to generate the refined caption c. The prompt is designed as: "The picture is stitched from the point cloud image and the RGB image of the same indoor object. There is a c^{env} near the object. Briefly describe the object in the picture." Through the refinement process, we obtain a precise object list $O = \{o_1, \ldots, o_N\}$, where each object o_i is represented as $o_i = \langle f_i, p_i, c_i \rangle$, consisting of semantic embedding f_i, pointcloud p_i, and precise node caption c_i.

3.3 Edge Caption Generation

Following [8], we estimate spatial relationships among 3D objects to complete the scene graph. For connectivity, we diverge from the fixed NN-ratio of [8] by introducing a dynamic threshold aligned with our downsampling strategy (Eq. (1)), adapting edge construction to varying point densities. For relationship captioning, although our paradigm inherently supports an open-vocabulary relation space, we intentionally constrain the final taxonomy to eight categories. These comprise bi-directional pairs for *support* (on), *containment* (in), and *part-whole* (part of), alongside *proximity* (near) and *none*. This multi-class constraint is strategically introduced to eliminate synonym ambiguity and ensure the stability of the LLM's spatial reasoning patterns. Moreover, we incorporate few-shot in-context examples. This design provides the model with explicit spatial reasoning patterns, ensuring that the generated edge captions are more expressive and accurate, thereby establishing a more robust structured representation for complex downstream robotics applications.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets & Benchmarks. We evaluate our method on two real-world datasets: **SceneFun3D** [4] (20 scenes) and **FunGraph3D** [43] (24 scenes), strictly following the OpenFunGraph [43] protocol. These datasets feature real-world challenges like sensor noise and clutter, inherently distinct from synthetic data. To align with standard 3DSSG definitions [35], we filter ground truths to retain only object nodes and semantic edges, excluding specialized functional interaction elements. Additionally, we use **Replica** [32] exclusively for fine-grained human evaluation, leveraging its clean geometry for unambiguous semantic assessment.

Baselines & Implementation Variants. We compare against SOTA open-vocabulary methods: **Open3DSG** [18], **ConceptGraphs** [8] (and its Detector variant), and **OpenFunGraph** [43]. To demonstrate scalability, we evaluate two framework variants: **Ours-LLaVA** (using a local LLaVA-v1.5-7b [22] for fair comparison) and **Ours-GPT** (using GPT-4o [11] to probe the performance upper bound). Additionally, three compact settings are included for mechanism analysis: (i) **Passive VLM** completely bypasses active re-shooting and aggregates captions solely from the top-3 highest-confidence passive crops. (ii) **Re-shot VLM** strips away the RAG module and image concatenation, directly adopting the isolated re-shot caption as the final node semantic. (iii) **Ours Node + CG Edge** pairs our rectified nodes with the vanilla ConceptGraphs [8] edge generator to isolate node-level contributions.

Implementation Details. We utilize SAM (sam_vit_h_4b8939) and CLIP (ViT-H-14) for segmentation and semantic extraction. Reasoning is powered by GPT-4o or LLaVA. Key hyperparameters include: similarity threshold $\delta_{sim} = 0.45$ (balancing over-segmentation and distinct instance discovery), base voxel size $\delta_{sample} = 0.01$m, and balanced scoring weights ($\alpha_{prior} = \alpha_{horiz} = 0.2, \alpha_{vis} = 0.6$) for uncertainty estimation. Exhaustive parameter analyses, prompt templates, and runtime profiling are deferred to the Supplementary Material.

Robustness Setting. To address concerns that 3D-centric pipelines over-rely on perfect multi-view reconstructions, we introduce an extreme stress test. We deliberately inject Gaussian noise ($\sigma_{trans} = 0.02\,\text{m}$, $\sigma_{rot} = 1^\circ$) into camera extrinsics. This *w/ Noisy Pose* setting (Tab. 1) empirically verifies if our re-shot and RAG mechanisms remain effective despite severely degraded point clouds.

Evaluation Metrics. For our main evaluation on SceneFun3D and FunGraph3D, we report Recall@K for both *Objects* (node classification) and *Edges* (predicate prediction). To ensure strictly fair comparisons and handle open-vocabulary edge predictions, we directly adopt the evaluation protocol of OpenFunGraph [43]: predicted predicates are mapped to ground-truth relationships based on the cosine similarity of their BERT embeddings, preventing any mapping bias from our expanded relation space. For the fine-grained Human Evaluation on Replica [32], three independent experts assess semantic correctness using a strict annotation protocol, with inter-annotator agreement verified via Fleiss' Kappa (κ).

Table 1: Quantitative comparison of 3DSG generation on real-world datasets. Following the standard evaluation protocol, we report Recall@K (R@K) for both Object classification (Nodes) and Relationship prediction (Edges) on the SceneFun3D (pink columns) and FunGraph3D (yellow columns). **Bold** and underline denote the best and second-best performances, respectively. * indicates the usage of OpenFunGraph's fused 3D nodes rather than the ground-truth for a fair comparison.

Type	Method	SceneFun3D				FunGraph3D			
		Objects		Edges		Objects		Edges	
		R@3	R@10	R@5	R@10	R@3	R@10	R@5	R@10
Baselines	Open3DSG [18]	61.2	70.7	69.2	78.8	47.6	55.7	47.9	55.9
	Open3DSG* [18]	42.9	50.0	64.4	72.3	30.9	44.1	46.6	55.7
	ConceptGraphs [8]	71.3	77.1	80.2	95.0	56.6	65.6	51.5	84.6
	ConceptGraphs (Det) [8]	72.3	78.6	81.3	92.0	57.5	68.0	52.4	80.2
	OpenFunGraph [43]	81.8	87.8	**88.1**	96.2	70.7	79.1	65.1	**91.4**
Analysis	Ours Node + CG Edge	–	–	84.1	95.6	–	–	61.2	89.4
	Passive VLM (gpt-4o)	67.7	72.9	78.3	86.3	53.4	66.1	49.5	78.2
	Passive VLM (gpt-5.2)	78.9	85.6	80.6	89.1	70.4	78.6	60.6	84.1
	Re-shot VLM (gpt-4o)	59.4	72.3	76.7	85.5	52.9	76.3	51.9	65.4
Ours	Ours-LLaVA	81.3	88.4	87.1	95.8	73.6	81.1	64.4	90.3
	w/ Noisy Pose	79.5	87.5	84.4	93.6	72.2	80.2	61.2	87.1
	Ours-GPT	**83.0**	**90.2**	86.2	**97.9**	**75.0**	**83.0**	**65.7**	**91.4**
	w/ Noisy Pose	82.1	89.3	86.7	95.7	73.6	82.1	62.5	87.3

4.2 Main Results

1) Object Classification (Node). Our method establishes a new state-of-the-art for object recognition across both datasets. As shown on the SceneFun3D split, **Ours-GPT** achieves a remarkable 90.2% R@10, outperforming the previous best method, OpenFunGraph (87.8%). This dominance extends to the more cluttered and challenging FunGraph3D benchmark, where **Ours-GPT** reaches 75.0% in R@3 and 83.0% in R@10, exceeding the SOTA by substantial margins of 4.3% and 3.9%, respectively. Crucially, even our locally deployed open-weight model (**Ours-LLaVA**) surpasses OpenFunGraph (73.6% vs. 70.7% in R@3). This compelling result underscores that our *diagnose-then-rectify* framework effectively compensates for raw model capacity, enabling accessible VLMs to achieve top-tier perception through active verification and RAG.

2) Relationship Prediction (Edge). Our method demonstrates superior performance on the challenging predicate prediction task. Although our edge generation paradigm resembles ConceptGraphs, our performance leap stems fundamentally from enhanced object node accuracy. Since relationship reasoning depends on anchor objects, our precise, less hallucinated node captions provide the LLM with superior context for inferring spatial interactions. This effectively

prevents cascading errors—where misclassified objects yield nonsensical relationships—that plague passive-observation baselines. Consequently, even without a heavy relational reasoning module, **Ours-GPT** achieves 97.9% R@10 on SceneFun3D (vs. OpenFunGraph's 96.2%) and 65.7% R@5 on FunGraph3D (vs. 65.1%). These gains confirm that resolving object-level semantic ambiguity is critical for improving overall scene graph connectivity.

3) Mechanism Analysis. We analyze the variants in Table 1 to examine the main components. **(a) Node Quality Drives Edge Gains:** *Ours Node + CG Edge* yields relationship predictions close to our full model (95.6% vs 97.9% R@10 on SceneFun3D) despite using the vanilla ConceptGraphs edge builder. This result indicates that the edge improvements are largely driven by enhanced object-node accuracy. **(b) Passive Observation Bottleneck:** *Passive VLM Baseline* (top-3 confident crops from passive tracks) underperforms our method by large margins. This shows that perception bottlenecks from viewpoint constraints cannot be resolved by simply scaling up backend foundation models. **(c) Failure of Isolated Re-shots:** Isolated re-shot captioning (*Re-shot VLM*) leads to drastic semantic degradation (dropping to 59.4% Object R@3 on SceneFun3D). This underscores that the tight synergy of our image concatenation and uncertainty-guided RAG is essential to prevent reshot degradation.

4) Robustness against Geometric Degradation. We stress-test our system by injecting camera pose noise during reconstruction to simulate tracking drift. As shown in Tab. 1 (*w/ Noisy Pose*), our framework exhibits exceptional resilience: the Object R@3 of **Ours-GPT** on SceneFun3D drops by merely 0.9% (83.0% to 82.1%), and **Ours-LLaVA** degrades by less than ∼3.2% on FunGraph3D. This disproves the concern that active re-shooting over-relies on perfect 3D reconstructions. This robustness stems from: (1) VLM's inherent perceptual tolerance to geometric ghosting when global semantic completeness is preserved; (2) our image concatenation strategy, which allows the VLM to fall back on the uncorrupted 2D crop as a dependable reference even under severe structural distortion.

4.3 Fine-grained Semantic Quality: Human Evaluation

While automated metrics assess recall against predefined ground truths, the open-vocabulary nature of our method inherently complicates automatic evaluation. Following ConceptGraphs [8], we resort to a fine-grained human evaluation to rigorously assess the semantic precision of the generated graphs. We conduct this qualitative audit on 8 scenes from the **Replica** [32], leveraging its high-fidelity geometry as an unambiguous reference, and compare our method against ConceptGraphs (CG) and ConceptGraphs-Detector (CG-D) [8].

Strict Annotation Protocol. To address the ambiguities inherent in open-vocabulary generation, we established a strict annotation protocol. We recruited three independent domain experts (students in 3D vision) to evaluate the randomized base units (point clouds, images, and predicted captions). Crucially, we defined a clear boundary for "accuracy": a node or edge caption is deemed *Valid*

Table 2: Human Evaluation on Replica [32] dataset. We evaluate fine-grained semantic quality across 8 indoor scenes. Blue columns denote Node-level metrics, and Green columns denote Edge-level metrics. **Prec.**: Precision; **Valid**: Count of valid objects; **Dup.**: Count of duplicate objects. **Bold** indicates best performance.

Scene	Node Quality (Objects)									Edge Quality (Relationships)					
	Precision ↑			Valid Count ↑			Duplicates ↓			Precision ↑			Valid Count ↑		
	Ours	CG	CG-D	Ours	CG	CG-D	Ours	CG	CG-D	Ours	CG	CG-D	Ours	CG	CG-D
room0	**0.87**	0.77	0.53	61	57	60	**1**	4	3	**0.93**	0.87	0.88	27	15	16
room1	**0.88**	0.73	0.71	51	45	42	**0**	5	3	**0.97**	0.92	0.91	30	12	11
room2	**0.85**	0.63	0.50	47	48	50	**0**	3	2	**0.94**	0.91	0.92	35	11	12
office0	**0.73**	0.61	0.61	48	44	41	**1**	1	1	**0.93**	0.78	0.82	27	9	11
office1	**0.73**	0.64	0.46	44	25	24	**0**	1	3	**0.93**	0.80	0.86	28	5	7
office2	**0.87**	0.77	0.68	67	48	44	**1**	3	2	**0.88**	0.79	0.86	34	14	14
office3	**0.85**	0.69	0.60	65	59	57	2	4	2	**0.84**	0.78	0.77	32	9	13
office4	**0.79**	0.61	0.57	53	41	46	**1**	5	4	**0.86**	0.67	0.80	22	3	5
Average	**0.82**	0.68	0.58	-	-	-	-	-	-	**0.91**	0.82	0.85	-	-	-

if it captures the correct semantic category and primary spatial/functional attributes. Acceptable paraphrases and synonymous descriptions (e.g., "a wooden table" vs. "a brown desk") are treated as equally valid semantic equivalents, provided they do not introduce hallucinated attributes or incorrect spatial logic.

Inter-Annotator Agreement (IAA). To ensure the reliability of the human labels, we conducted a consistency check among the three annotators. We computed the Fleiss' Kappa (κ) score across a subset of 100 randomly sampled evaluation units. The annotators achieved a κ score of 0.76, indicating "substantial agreement." Disagreements primarily stemmed from ambiguous linguistic expressions or subjective interpretations of spatial relationships, which were subsequently resolved via majority voting for the final tally.

Results Analysis. Table 2 reports the comprehensive evaluation across node and edge qualities. Our method consistently outperforms both ConceptGraphs (CG) and ConceptGraphsDetector (CG-D) across most evaluation metrics. In terms of node precision, our method achieves an average score of 0.82, which is notably higher than CG (0.68) and CG-D (0.58), demonstrating the effectiveness of our re-shot guided uncertainty estimation in reducing noise during object caption aggregation. For edge precision, our method also achieves the highest average score (0.91), surpassing CG (0.82) and CG-D (0.85), indicating that our structured prompt and refined relationship categories lead to more accurate and interpretable edge captions. In addition, our method substantially reduces duplicate predictions while maintaining a higher number of valid objects and edges, further confirming its robustness.

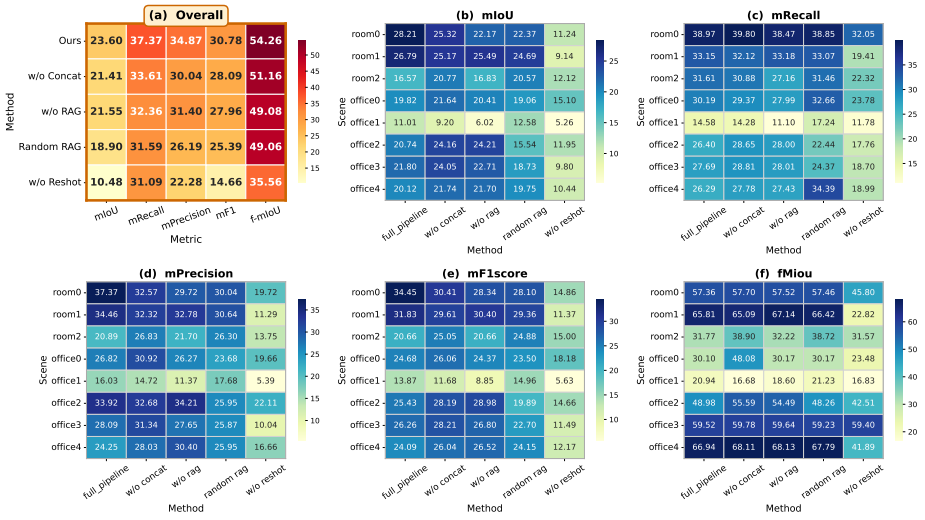

Fig. 3: Ablation study quantifying component contributions. Performance degradation is observed when removing reshot guided uncertainty estimation (w/o Reshot), RAG (w/o RAG), concatenated re-shot images (w/o Concat), or using random retrieval (Random RAG), confirming the complementary roles of all proposed components.

4.4 Ablation Study

We conduct ablation studies to quantify the contribution of each component in our pipeline. Our full model (*Ours*) consists of dynamic downsample-mapping & fusion, re-shot guided uncertainty estimation, and node-level RAG with concatenated re-shot image prompts. We compare against several variants: (i) removing re-shot guided uncertainty estimation (*w/o Reshot*), (ii) removing RAG (*w/o RAG*), (iii) applying random retrieval for RAG (*random-RAG*), and (iv) removing the concatenated re-shot image prompts (*w/o Concat*).

Experiments are conducted on the Replica dataset [32]. Following the protocol used in [8, 12], we obtain per-point semantic GT labels by fusing 2D masks into 3D via gradSLAM [13]. To ensure faithful open-vocabulary evaluation, we use GPT-4o as a semantic assigner to match GT labels with predicted node captions. After 1-NN point-cloud matching and confusion matrix construction, quantitative results are reported in Fig. 3, where (a) shows overall performance and (b)-(f) detail metric-specific heatmaps across individual scenes. Ablation results confirm that removing Reshot, RAG, Concat, or using Random RAG leads to performance drops, highlighting their complementary roles.

As shown in Fig. 3 (a), removing any component leads to a performance drop, confirming their complementary roles. Notably, removing re-shot guided uncertainty estimation (*w/o Reshot*) causes the most drastic degradation, e.g., in mF1 (14.66 vs. 30.78) and f-mIoU (35.56 vs. 54.26). This is because *Reshot* acts as a fundamental filter; without it, the system is forced to fuse unreliable or hallu-

cinated captions, which severely degrades the global semantic map. Eliminating RAG (*w/o RAG*) reduces both precision and overall accuracy, while replacing it with random retrieval (*random-RAG*) further deteriorates performance, highlighting the necessity of reliable retrieval. Removing the concatenated re-shot image (*w/o Concat*) also leads to lower recall and f-mIoU, suggesting that multi-view prompts alleviate viewpoint bias and enrich object descriptions. These results collectively demonstrate that all three proposed components contribute significantly to the robustness and accuracy of our framework.

4.5 Real-World Robotic Deployment

To further validate the practical viability of our approach, we deploy RAG-3DSG on a mobile robot and conduct experiments in a real-world indoor facility.

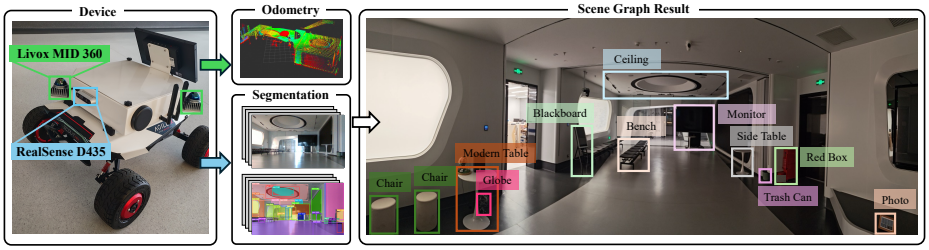

Fig. 4: Real-world robotic deployment. By fusing RGB-D perception with LiDAR-IMU odometry, our mobile platform successfully drives the RAG-3DSG framework to generate high-fidelity semantic 3D scene graphs in real-world indoor environments.

Hardware Setup. As illustrated in Fig. 4, our custom robotic platform is built upon an Agilex SCOUT MINI chassis. The sensor suite comprises a RealSense D435 RGB-D camera for visual perception, a CH110 IMU, and dual Livox MID 360 LiDARs to provide omnidirectional geometric coverage and blind-spot reduction. All data processing and sensor synchronization are handled by an onboard NVIDIA Jetson AGX Xavier computer.

Deployment Pipeline. During the physical deployment, the robot teleoperates through the facility, continuously capturing RGB and Depth streams. Simultaneously, we maintain a LiDAR-IMU Odometry (LIO) pipeline to estimate real-time camera poses. These poses are strictly utilized to perform cross-frame matching and tracking of 2D segmented object instances. Once the data collection is complete, the fused 3D object observations are fed into our RAG-3DSG framework to generate the final textual 3D semantic scene graph.

Results and Robustness Analysis. Aligning with our stress tests, physical deployment reveals the system's profound resilience. Under inferior localization (e.g., drift-prone pure IMU), corrupted point clouds degrade the active *re-shot* mechanism via ghosting. Yet, the system only degrades marginally, leveraging the

RAG module's semantic anchors for robust contextual refinement. Conversely, integrating robust state-estimation (e.g., Fast-LIO2 [38]) unleashes the full synergy of our re-shot and RAG modules, yielding high-fidelity scene graphs.

4.6 Runtime Analysis and System Scalability

To address the computational overhead nature of foundation models, we strictly optimize our framework for practical scalability. While currently operating as an offline system, our dynamic downsample-mapping strategy significantly accelerates the initial 3D fusion phase by pruning pointcloud, reducing the per-iteration processing time by nearly two-thirds compared to ConceptGraphs [8] (e.g., from 6.65s to 2.49s on Replica). For semantic reasoning, the computational and monetary costs remain highly manageable. Our LLaVA variant runs efficiently on a single local consumer GPU, while the GPT-4o variant incurs a minimal API cost of ∼$0.50 per scene on FunGraph3D. Under the same backbone as ConceptGraphs (VLM reasoning costs of ∼0.78s/iteration), the additional active verification introduces a rendering cost of ∼0.26s/object. Finally, although designed for offline high-precision generation, our pipeline aligns well with modern decoupled robotic architectures (e.g., DAAAM [7]). This paradigm allows our method to be seamlessly integrated as a low-frequency semantic backend to guide high-frequency real-time exploration in future online applications.

5 CONCLUSION

In this work, we present RAG-3DSG for generating accurate and robust open-vocabulary 3D scene graphs. We are the first to address semantic inconsistency in cross-view aggregation by introducing active re-shot verification and an object-level RAG module for contextual refinement. Extensive evaluations on real-world datasets demonstrate new state-of-the-art performance in both node and edge accuracy, while human evaluation confirms the generation of cleaner, less hallucinated graphs. Crucially, extreme geometric stress tests and physical robotic deployment validate our system's exceptional resilience against sensor noise, paving the way for reliable 3D perception in complex downstream tasks.

Acknowledgements

Sihong Xie was supported by the National Key R&D Program of China (Grant No. 2023YFF0725001), the Department of Science and Technology of Guangdong Province (2023CX10X079), the Guangzhou-HKUST (GZ) Joint Funding Program (Grant No. 2023A03J0008), and the Education Bureau Guangzhou Municipality, to which the authors extend their sincere gratitude for the support.

References

1. Agia, C., Jatavallabhula, K.M., Khodeir, M., Miksik, O., Vineet, V., Mukadam, M., Paull, L., Shkurti, F.: Taskography: Evaluating robot task planning over large 3d scene graphs. In: Conference on Robot Learning. pp. 46–58. PMLR (2022)
2. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5664–5673 (2019)
3. Chen, R., Chang, Y., Tang, X., Chen, H., Xie, S.: Psg-nav: Probabilistic scene graph navigation via multiverse decision making. arXiv preprint arXiv:2606.01313 (2026)
4. Delitzas, A., Takmaz, A., Tombari, F., Sumner, R., Pollefeys, M., Engelmann, F.: Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14531–14542 (2024)
5. Gadre, S.Y., Wortsman, M., Ilharco, G., Schmidt, L., Song, S.: Clip on wheels: Zero-shot object navigation as object localization and exploration. arXiv preprint arXiv:2203.10421 **3**(4), 7 (2022)
6. Gay, P., Stuart, J., Del Bue, A.: Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. In: Asian Conference on Computer Vision. pp. 330–346. Springer (2018)
7. Gorlo, N., Schmid, L., Carlone, L.: Describe anything anywhere at any moment. arXiv preprint arXiv:2512.00565 (2025)
8. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
9. Honerkamp, D., Büchner, M., Despinoy, F., Welschehold, T., Valada, A.: Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. IEEE Robotics and Automation Letters (2024)
10. Hughes, N., Chang, Y., Carlone, L.: Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. arXiv preprint arXiv:2201.13360 (2022)
11. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
12. Jatavallabhula, K.M., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Maalouf, A., Li, S., Iyer, G., Saryazdi, S., Keetha, N., et al.: Conceptfusion: Open-set multimodal 3d mapping. arXiv preprint arXiv:2302.07241 (2023)
13. Jatavallabhula, K.M., Saryazdi, S., Iyer, G., Paull, L.: gradslam: Automagically differentiable slam. arXiv preprint arXiv:1910.10672 (2019)
14. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3668–3678 (2015)
15. Ju, Y., Liang, Y., Wang, Y.J., Gireesh, N., Ju, Y., Lee, S., Gu, Q., Hsieh, E., Huang, F., Sreenath, K.: Momagraph: State-aware unified scene graphs with vision-language model for embodied task planning. arXiv preprint arXiv:2512.16909 (2025)

16. Kim, U.H., Park, J.M., Song, T.J., Kim, J.H.: 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. IEEE transactions on cybernetics **50**(12), 4921–4933 (2019)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
18. Koch, S., Vaskevicius, N., Colosi, M., Hermosilla, P., Ropinski, T.: Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14183–14193 (2024)
19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
20. Li, H., Zhu, G., Zhang, L., Jiang, Y., Dang, Y., Hou, H., Shen, P., Zhao, X., Shah, S.A.A., Bennamoun, M.: Scene graph generation: A comprehensive survey. Neurocomputing **566**, 127052 (2024)
21. Linok, S., Zemskova, T., Ladanova, S., Titkov, R., Yudin, D., Monastyrny, M., Valenkov, A.: Beyond bare queries: Open-vocabulary object grounding with 3d scene graph. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 13582–13589. IEEE (2025)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
23. Liu, H., Yan, N., Mortazavi, M., Bhanu, B.: Fully convolutional scene graph generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11546–11556 (2021)
24. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European conference on computer vision. pp. 852–869. Springer (2016)
25. Maggio, D., Chang, Y., Hughes, N., Trang, M., Griffith, D., Dougherty, C., Cristofalo, E., Schmid, L., Carlone, L.: Clio: Real-time task-driven open-set 3d scene graphs. IEEE Robotics and Automation Letters (2024)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., Suenderhauf, N.: Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. arXiv preprint arXiv:2307.06135 (2023)
28. Rashid, A., Sharma, S., Kim, C.M., Kerr, J., Chen, L.Y., Kanazawa, A., Goldberg, K.: Language embedded radiance fields for zero-shot task-oriented grasping. In: 7th Annual Conference on Robot Learning (2023)
29. Rosinol, A., Violette, A., Abate, M., Hughes, N., Chang, Y., Shi, J., Gupta, A., Carlone, L.: Kimera: From slam to spatial perception with 3d dynamic scene graphs. The International Journal of Robotics Research **40**(12-14), 1510–1546 (2021)
30. Shah, D., Osiński, B., Levine, S., et al.: Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Conference on robot learning. pp. 492–504. PMLR (2023)
31. Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation. In: Conference on robot learning. pp. 894–906. PMLR (2022)

32. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
33. Sun, S., Zhi, S., Heikkilä, J., Liu, L.: Evidential uncertainty and diversity guided active learning for scene graph generation. In: The Eleventh International Conference on Learning Representations (2023)
34. Vázquez, P.P., Feixas, M., Sbert, M., Heidrich, W.: Viewpoint selection using viewpoint entropy. In: VMV. vol. 1, pp. 273–280 (2001)
35. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3961–3970 (2020)
36. Werby, A., Huang, C., Büchner, M., Valada, A., Burgard, W.: 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 (2024)
37. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7515–7525 (2021)
38. Xu, W., Cai, Y., He, D., Lin, J., Zhang, F.: Fast-lio2: Fast direct lidar-inertial odometry. IEEE Transactions on Robotics **38**(4), 2053–2073 (2022)
39. Yamazaki, K., Hanyu, T., Vo, K., Pham, T., Tran, M., Doretto, G., Nguyen, A., Le, N.: Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 9411–9417. IEEE (2024)
40. Yan, Z., Li, S., Wang, Z., Wu, L., Wang, H., Zhu, J., Chen, L., Liu, J.: Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation. IEEE Robotics and Automation Letters (2025)
41. Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., Loy, C.C.: Zoom-net: Mining deep feature interactions for visual relationship recognition. In: Proceedings of the European conference on computer vision (ECCV). pp. 322–338 (2018)
42. Yin, H., Xu, X., Wu, Z., Zhou, J., Lu, J.: Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. Advances in neural information processing systems **37**, 5285–5307 (2024)
43. Zhang, C., Delitzas, A., Wang, F., Zhang, R., Ji, X., Pollefeys, M., Engelmann, F.: Open-vocabulary functional 3d scene graphs for real-world indoor spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19401–19413 (2025)
44. Zhang, Z., Wu, Y., Jia, L., Wang, Y., Zhang, Z., Li, Y., Ran, B., Zhang, F., Sun, Z., Yin, Z., et al.: Think3d: Thinking with space for spatial reasoning. arXiv preprint arXiv:2601.13029 (2026)
45. Zhao, H., Liu, A., Zhang, Z., Wang, W., Chen, F., Zhu, R., Haffari, G., Zhuang, B.: Cov: Chain-of-view prompting for spatial reasoning. arXiv preprint arXiv:2601.05172 (2026)
46. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in neural information processing systems **36**, 19769–19782 (2023)