Open-World 3D Scene Graph Generation for Retrieval-Augmented Reasoning

Yu Fei[1], **Quan Deng**[2], **Shengeng Tang**[3], **Li Yuehua**[4], **Lechao Cheng**[3*]

[1] Liaoning University of Technology
[2] University of the Chinese Academy of Sciences
[3] Hefei University of Technology
[4] Zhejiang Lab

Abstract

Open-world 3D scene understanding is fundamentally challenging for vision and robotics, due to the constraints of closed-vocabulary supervision and static annotations. To address this, we propose a unified framework for Open-World 3D Scene Graph Generation with Retrieval-Augmented Reasoning, which enables generalizable and interactive 3D scene understanding. Our method integrates vision-language models with retrieval-based reasoning to support multimodal exploration and language-guided interaction. The framework comprises two key components: (1) a dynamic scene graph generation module that detects objects and infers semantic relationships without fixed label sets, and (2) a retrieval-augmented reasoning pipeline that encodes scene graphs into a vector database to support text/image-conditioned queries. We evaluate our method on 3DSSG and Replica benchmarks across four tasks—scene question answering, visual grounding, instance retrieval, and task planning—demonstrating robust generalization and superior performance in diverse environments. Our results highlight the effectiveness of combining open-vocabulary perception with retrieval-based reasoning for scalable 3D scene understanding [1].

Introduction

Understanding 3D scenes is fundamental for tasks like autonomous navigation and augmented reality. However, most traditional methods rely on closed-vocabulary annotations and curated datasets, which hinders scalability in dynamic, unstructured settings—such as 2D dynamic video scene understanding (Yu, Ji, and Li 2025)—where novel objects and interactions continually emerge. This gap calls for open-world 3D understanding methods that can generalize to unseen scenarios without exhaustive supervision.

Recent advances in vision-language models (VLMs) such as CLIP (Hafner et al. 2021), ALIGN (Jia et al. 2021), and ImageBind (Girdhar et al. 2023) have enabled open-vocabulary learning, which shows remarkable potential in 2D tasks like classification and detection. These successes have inspired efforts to adapt VLMs for 3D vision, leading to progress in tasks such as semantic segmentation and object recognition. However, these methods often rely on 2D-3D projections and RGB-D inputs with known poses, which are impractical in many real-world settings due to sensor limitations, occlusions, and viewpoint variations.

To address these limitations, we introduce a retrieval-augmented open-world 3D scene graph generation framework. It consists of two core components: (1) a dynamic scene graph module that constructs semantic and spatial representations of the environment by identifying objects and relationships without fixed label sets; and (2) a retrieval-augmented navigation system that enables natural language-based exploration of 3D scenes.

This work is motivated by the need for adaptive and scalable 3D understanding systems that can operate in open-world settings without extensive human supervision. While conventional methods perform well in constrained environments, they struggle to generalize to dynamic and diverse scenarios. By integrating VLMs with retrieval mechanisms, our method enables compositional reasoning over unseen objects and relationships. Our framework addresses the following challenges: (1) **Open-Vocabulary Grounding**: Recognizing and localizing unseen objects and relations without predefined labels. (2) **Dynamic Graph Construction**: Building and updating 3D scene graphs that capture evolving semantic and spatial relationships. (3) **Language-Guided Interaction**: Supporting navigation and querying through natural language, requiring alignment between textual and spatial modalities.

Our contributions are summarized as follows:

- A novel framework for open-world 3D scene graph generation, leveraging vision-language models to extract objects and relationships without fixed annotations.
- A retrieval-augmented reasoning module that enables flexible scene interaction through query-based exploration.
- Extensive experiments on 3DSSG and Replica benchmarks demonstrating improved performance and generalization in dynamic, real-world environments.

Related Work

Open-Vocabulary Scene Understanding. The success of 2D vision-language models (VLMs) like CLIP (Hafner et al.

*Correspondding Author: chenglc@hfut.edu.cn

[1]Our code is available at https://anonymous.4open.science/r/3DSU-B04D.

2021), ALIGN (Jia et al. 2021), and ImageBind (Girdhar et al. 2023) has driven efforts to extend open-vocabulary recognition (Wang et al. 2024a; Tian et al. 2024) to 3D tasks, including semantic and instance segmentation (Ha and Song 2022; Hegde, Valanarasu, and Patel 2023; Zhang et al. 2022). Existing methods often fuse CLIP-based supervision with 3D backbones using posed RGB-D data (Tang et al. 2022; Wu, Hong, and Tang 2024), yet remain constrained by projection errors, occlusion artifacts, and reliance on pose priors.

Open-Vocabulary Scene Graph Generation. 3D scene graph generation models semantic and spatial relationships among objects, linking visual perception with symbolic reasoning. Prior works (Wald et al. 2020; Ren et al. 2023) focused on supervised indoor data, while recent approaches such as Open3DSG (Koch et al. 2024) and zero-shot variants (Linok et al. 2024) employ VLMs to recognize unseen objects and infer spatial relations. However, many still depend on annotated RGB-D inputs or fixed camera poses, limiting open-world adaptability. In contrast, our method eliminates the need for manual annotations or fixed-pose RGB-D data, enabling fully annotation-free 3D scene graph generation. It achieves strong performance in both open- and closed-vocabulary benchmarks while generalizing better to unseen environments.

Efficient Multimodal Large Language Models Adaptation. Multimodal Large Language Models (MLLMs) enable joint reasoning over visual and textual inputs (Wu et al. 2023a; Hu et al. 2023; Che et al. 2024; Wang et al. 2024b; Zhang et al. 2025; Wang et al. 2025; Li et al. 2025; Shen et al. 2025). They typically combine an LLM backbone (e.g., LLaMA (Dubey et al. 2024), Qwen (Bai et al. 2023)), a visual encoder (e.g., CLIP (Radford et al. 2021), BLIP (Li et al. 2022)), and adapter modules aligning visual features with the LLM's embedding space (Yin et al. 2023). Given their scale, training MLLMs from scratch is costly (Jiang et al. 2025). Parameter-Efficient Fine-Tuning (PEFT) techniques such as LoRA (Hu et al. 2022) and Prefix Tuning (Li and Liang 2021) reduce memory and computation by updating only a small parameter subset. Nevertheless, challenges persist in adapting MLLMs to long-tail or domain-specific scenarios requiring fine-grained cross-modal alignment, underscoring the need for improved adaptation strategies that maintain generalization while enhancing task-specific performance (Hu et al. 2022; Li and Liang 2021; He et al. 2025).

Methodology

To enable robust navigation in open-world environments, our approach builds on retrieval-augmented reasoning over 3D scene graphs, which encode object instances, spatial relations, and high-level semantics. Unlike closed-world settings with fully known and static layouts, open-world scenarios introduce challenges such as dynamic object placement, occlusion, and real-time scene changes. These issues require scene representations that are both adaptive and incrementally updatable. To address this, we propose a retrieval-augmented framework that leverages previously stored scene graphs to support inference and decision-making under partial observations. By retrieving relevant spatial-semantic context from past interactions, the agent can recover missing structures, disambiguate uncertain observations, and guide navigation in dynamically evolving environments.

Definition 1: A ***3D scene graph*** provides a structured graph-based representation of a 3D environment. Formally, a 3D scene graph is defined as a directed graph:

$$G_{3D} = (O, \mathcal{R}), \tag{1}$$

where O is the set of *objects*, each represented as a node; $\mathcal{R}$ is the set of directed edges, where each edge encodes a relational dependency between two objects.

Each object $o_i \in O$ is characterized by the attributes:

- A category label $l_i \in \mathcal{L}$, where $\mathcal{L}$ is the set of possible semantic categories.
- An optimal viewpoint $T^{c*}_{w,i} \in SE(3)$ maximizes visibility and projected 2D coverage of object o_i, representing the camera pose in the world frame[2].
- A feature descriptor $f_i \in \mathbb{R}^{D_f}$, where $D_f \in \mathbb{N}^+$ is the fixed feature dimension.
- An oriented bounding box (OBB) $b_i = (c_i, \ell_i, R_i)$ defined as:

$$c_i \in \mathbb{R}^3, \quad \ell_i \in \mathbb{R}^3_{>0}, \quad R_i \in \mathrm{SO}(3). \tag{2}$$

  Here, c_i is the bounding box centroid, ℓ_i represents the size (length, width, height), and R_i is the rotation matrix[3].
- A node category $c_{\text{node},i} \in \mathcal{C}_{node}$, specifying its role in the scene.

Each directed edge $\mathcal{R}_{ij} \in \mathcal{R}$ encodes a relationship between o_i and o_j, represented as:

$$\mathcal{R}_{ij} = (o_i, r_{ij}, o_j), \quad r_{ij} \in \mathcal{C}_{edge}. \tag{3}$$

where r_{ij} represents the inferred semantic relationship.

This structured representation explicitly models object properties, spatial configurations, and semantic interactions, facilitating robust open-world reasoning and robotic decision-making in dynamic 3D environments.

Overview. We present a unified framework for Open-World 3D Scene Graph Generation with Retrieval-Augmented Reasoning (see Figure 1), enabling multimodal understanding and interaction in 3D environments. From multi-frame RGB-D input, we construct object-centric scene graphs via pose-aware detection, best-view selection, and VLM-based relation extraction. These graphs are encoded into a vector database for efficient semantic retrieval. Given text or image queries, our system composes grounded prompts for LLMs

[2]$\mathrm{SE}(3) = \{(R, t) \mid R \in \mathrm{SO}(3),\ t \in \mathbb{R}^3\}$: 3D rigid body transformations.

[3]$\mathrm{SO}(3) = \{R \in \mathbb{R}^{3\times3} \mid R^\top R = I,\ \det(R) = +1\}$: group of 3D rotation matrices; $\mathbb{R}^3_{>0}$: 3D vector with strictly positive components.

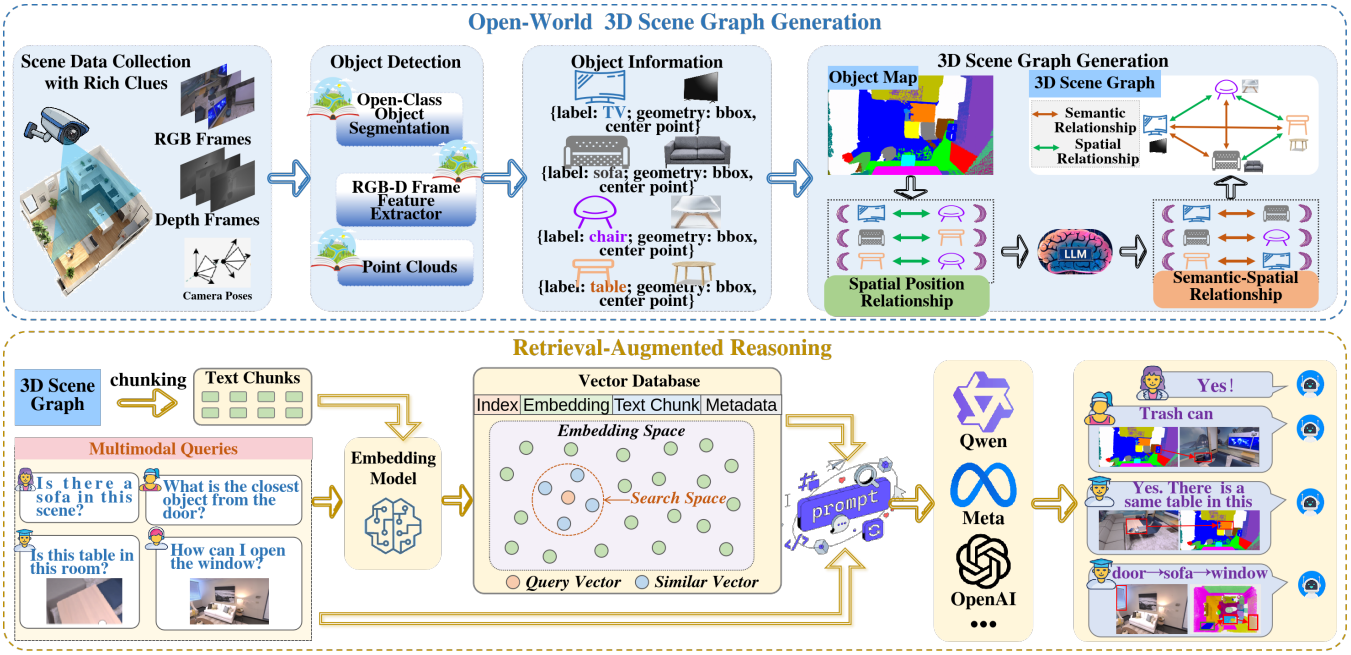

Figure 1: Overview of the proposed framework for Open-World 3D Scene Graph Generation and Retrieval-Augmented Navigation. The framework comprises two key components: (1) a 3D Scene Graph Generator that incrementally builds semantic and spatial representations from RGB-D sequences by detecting objects, estimating poses, selecting optimal viewpoints, and extracting inter-object relations via vision-language reasoning; and (2) a Retrieval-Augmented Reasoning module that transforms the scene graph into a vectorized knowledge base to support three categories of interaction: (i) spatial object queries, (ii) semantic relationship reasoning, and (iii) instance-level retrieval. This integrated design enables grounded, multimodal, and context-aware interaction within dynamic open-world 3D environments.

to support four tasks: question answering, visual grounding, instance retrieval, and task planning—facilitating open-vocabulary, context-aware reasoning for embodied applications. We now delve into the technical details of our framework, covering the construction of open-world 3D scene graphs, the retrieval-augmented reasoning pipeline, and its application across four multimodal interaction tasks.

Open-World 3D Scene Graph Generation

Multi-Frame Object Detection. Given a sequence of RGB-D frames, the open-scene 3D object detector identifies and represents objects within a 3D scene. Each frame consists of a color image $I \in \mathbb{R}^{H\times W\times 3}$, a depth map $D \in \mathbb{R}^{H\times W}$, intrinsic camera parameters $K \in \mathbb{R}^{3\times 3}$, and a camera pose represented as a rigid body transformation in SE(3):

$$T_w^c = \begin{bmatrix} R_w^c & t_w^c \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathrm{SE}(3), \tag{4}$$

where $R_w^c \in \mathrm{SO}(3)$ is the rotation matrix and $t_w^c \in \mathbb{R}^3$ is the translation vector. The inverse transformation to convert a 3D point from the camera frame to the world frame is:

$$T_c^w = (T_w^c)^{-1} = \begin{bmatrix} (R_w^c)^\top & -(R_w^c)^\top t_w^c \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{5}$$

To account for pose uncertainty, we model Gaussian noise in the Lie Algebra $\mathfrak{se}(3)$:

$$\mathcal{T}_w^c = \log(T_w^c) \sim \mathcal{N}(\hat{\mathcal{T}}_w^c, \Sigma_{\text{pose}}), \tag{6}$$

where $\mathcal{T}_w^c = \begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\omega} \end{bmatrix} \in \mathbb{R}^6$ consists of the translation perturbation $\boldsymbol{\xi}$ and rotation perturbation $\boldsymbol{\omega}$. The perturbed pose is reconstructed as:

$$\tilde{T}_w^c = \exp(\mathcal{T}_w^c) \cdot T_w^c. \tag{7}$$

Each detected object o_i is represented by an oriented 3D bounding box:

$$b_i = (c_i, \ell_i, R_i), \quad c_i \in \mathbb{R}^3, \quad \ell_i \in \mathbb{R}^3_{>0}, \quad R_i \in \mathrm{SO}(3), \tag{8}$$

where c_i is the bounding box centroid in world coordinates, ℓ_i represents its dimensions (length, width, height), and R_i is a rotation matrix defining its orientation. Each object detection is assigned a confidence score $\sigma_i$modeled by a Beta distribution:

$$\sigma_i \sim \mathrm{Beta}(\alpha_i, \beta_i), \quad \alpha_i = \mu_i \tau, \quad \beta_i = (1-\mu_i)\tau, \tag{9}$$

where μ_i is the predicted probability and τ is an adaptive scaling factor:

$$\tau = \max(\tau_{\min}, \min(\tau_{\max}, \tau_0 + \lambda \cdot \mathrm{entropy}(\mu_i))). \tag{10}$$

Each object mask $\mathcal{M}_i$ is used to extract 3D points via back-projection:

$$X_j^c = K^{-1}\,[u_j\ v_j\ 1]\,D_j, \quad \forall (u_j, v_j) \in \mathcal{M}_i, \tag{11}$$

where (u_j, v_j) are pixel coordinates and D_j is the depth value. The corresponding world coordinates are:

$$X_j^w = T_c^w X_j^c. \tag{12}$$

For feature consistency, objects are merged every L frames based on cosine similarity:

$$S(\tilde{f}_i, \tilde{f}_j) = \begin{cases} \frac{\langle \tilde{f}_i, \tilde{f}_j \rangle}{\|\tilde{f}_i\| \|\tilde{f}_j\|}, & \text{if } \frac{\|\tilde{f}_i - \tilde{f}_j\|}{\|\tilde{f}_i\| + \|\tilde{f}_j\|} < \tau_{\text{merge}} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

where $\tilde{f}_i$ is Mahalanobis-whitened:

$$\tilde{f}_i = \Sigma^{-1/2}(f_i - \mu_f). \tag{14}$$

Best-View Selection and Labeling. For viewpoint selection, the optimal viewpoint maximizes visibility and projected object coverage:

$$T_{w,i}^{c*} = \arg\max_{T_w^c \in \mathcal{P}} \left[A(\mathcal{P}(X_i^w, T_w^c)) \cdot V(X_i^w, T_w^c)^\gamma - \lambda D(T_{w,i}^c, T_w^c) \right]. \tag{15}$$

where $A(\mathcal{P}(X_i^w, T_w^c))$ is the projected object area, $V(X_i^w, T_w^c)$ is the visibility function, and $D(T_{w,i}^c, T_w^c)$ penalizes large pose changes:

$$D(T_{w,i}^c, T_w^c) = \|R_{w,i}^c - R_w^c\|_F + \|t_{w,i}^c - t_w^c\|_2. \tag{16}$$

Finally, each detected object is labeled by LLaVA (Li et al. 2024) using its best-view segmentation mask.

Reliable Object Filtering. To extract spatially meaningful object pairs, we compute the Euclidean distance and 3D Intersection over Union (IoU) between each pair of detected objects. The Euclidean distance is given by:

$$d_{ij} = \|c_i - c_j\|_2 = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}, \tag{17}$$

where $c_i, c_j \in \mathbb{R}^3$ denote the centroids of 3D bounding boxes b_i and b_j. The IoU measures volumetric overlap:

$$\text{IoU}(b_i, b_j) = \frac{\text{Vol}(b_i \cap b_j)}{\text{Vol}(b_i \cup b_j)}. \tag{18}$$

We retain object pairs satisfying:

$$\text{IoU}(b_i, b_j) > 0 \quad \text{or} \quad d_{ij} < d_{\text{thresh}}, \tag{19}$$

with $d_{\text{thresh}} = 0.5$ meters empirically chosen to capture relevant spatial interactions.

Semantic Relationship Extraction. To refine object-level interactions, we employ a vision-language model (Qwen2-VL-72B-Instruct-GPTQ-Int4) to infer the top-5 semantic predicates for each spatially valid object pair (o_i, o_j). These predicates characterize inter-object relations and are structured into a 3D semantic scene graph (3DSG):

$$\mathcal{R}_{ij} = (o_i, r_{ij}, o_j), \tag{20}$$

where $r_{ij} \in \mathcal{C}_{edge}$ denotes the predicted semantic relation. Each node o_i includes:

$$o_i = (l_i, f_i, b_i, T_{w,i}^{c*}), \tag{21}$$

with category label l_i, feature descriptor f_i, 3D bounding box b_i, and best-view pose $T_{w,i}^{c*}$.

To ensure task relevance, background elements (e.g., floor, ceiling) are filtered out. The resulting 3DSG captures both spatial and semantic contexts, facilitating downstream tasks such as embodied navigation and manipulation.

Retrieval-Augmented Semantic Reasoning

To support multimodal QA and navigation, we transform the 3DSG into a vector-searchable knowledge structure.

Vector Database Construction. To facilitate semantic indexing, we reorganize the 3D scene graph into structured representations termed *chunks*. Each chunk $\boldsymbol{\eta}_i$ centers on a specific object label (e.g., "*book*") and aggregates all corresponding instances in the scene, including their 3D attributes such as bounding boxes, optimal viewpoints, and textual descriptions, as well as their spatial and semantic relationships with other objects. This compact representation captures both intra-class diversity and inter-object contextual cues.

Each chunk $\boldsymbol{\eta}_i$ is then projected into a high-dimensional vector space via a semantic encoder ϕ, implemented using a pretrained language or vision-language model such as CLIP, BERT, or Text2Vec:

$$\boldsymbol{\zeta}_i = \phi(\boldsymbol{\eta}_i), \quad \boldsymbol{\zeta}_i \in \mathbb{R}^d. \tag{22}$$

The embedding preserves semantic consistency, mapping semantically related objects and relationships to nearby positions in the latent space, thereby enabling robust similarity-based retrieval.

The resulting embeddings and their associated chunks are stored in a vector database:

$$\mathcal{D} = \{(\boldsymbol{\zeta}_i, \boldsymbol{\eta}_i)\}_{i=1}^N, \tag{23}$$

where N denotes the number of unique label-centered chunks. This structure supports scalable and efficient retrieval of semantically relevant scene components for downstream reasoning tasks.

Grounded Prompt-Based Reasoning. Given a user query q in either text or image form, we first encode it into a semantic embedding $\boldsymbol{\xi}_q = \phi(q)$ using the same encoder ϕ employed during vector database construction. A top-k similarity search is then conducted over the database $\mathcal{D}$ to retrieve the most relevant scene chunks:

$$\mathcal{E}_q = \text{Top-}k(\mathcal{D}, \boldsymbol{\xi}_q), \tag{24}$$

where similarity is computed via cosine distance:

$$\cos(\boldsymbol{\xi}_q, \boldsymbol{\zeta}_i) = \frac{\boldsymbol{\xi}_q \cdot \boldsymbol{\zeta}_i}{\|\boldsymbol{\xi}_q\| \|\boldsymbol{\zeta}_i\|}. \tag{25}$$

Each retrieved chunk is then decomposed into structured components including object-level attributes Γ_i (e.g., category, viewpoint, position) and relationship information Λ_i with other objects:

$$\mathcal{E}_q = \bigcup_{i=1}^k (o_i, \Gamma_i, \Lambda_i). \tag{26}$$

These multimodal elements are integrated with the user query into a structured prompt, which is fed into a large language model (LLM) for grounded reasoning. For example, given a question such as:

"What is to the left of the chair?"

and retrieved facts such as *"chair at center, table on the left"*, the prompt becomes:

"Based on the information from the image: chair at center, table on left, please answer the question: What is to the left of the chair?"

The LLM (e.g., Qwen-2-72B-Instruct) processes this prompt by analyzing spatial configurations and object relationships. For instance, given the question:

"What is the largest object in front of the sofa?"

the model performs grounded reasoning over the graph-aware scene context to execute:

$$\text{argmax}_{o_i} \, \text{size}(o_i), \quad \forall o_i \in \text{in front of sofa}.$$

By leveraging both visual object features (e.g., bounding box dimensions, best-view attributes) and spatial relations (e.g., relative distance, orientation, containment), the system produces coherent and context-aware responses. For the above example, the answer may be:

"The table is in front of the sofa and is the largest object based on its projected area."

This grounded reasoning capability allows the system to handle complex queries involving attribute comparisons (e.g., size, color), spatial understanding (e.g., adjacency, alignment), and semantic relationship inference (e.g., on, beside, in front of), making it suitable for open-world 3D-VQA and embodied interaction tasks.

Scene-Driven Multimodal Interaction

We design a unified framework to support four types of scene-aware interaction tasks that bridge language, vision, and 3D semantics. All tasks are modeled as a function:

$$\mathcal{F}_{\text{int}} : \mathcal{I}_{\text{in}} \rightarrow \mathcal{I}_{\text{out}}, \tag{27}$$

where $\mathcal{I}_{\text{in}}$ denotes the input modality (text, image, or both), and $\mathcal{I}_{\text{out}}$ denotes the system output (textual answer, visual grounding, instance retrieval result, or action plan). Each task builds upon the vector-indexed 3D scene graph and the retrieval-augmented reasoning module.

Task I: Text-Based Scene Question Answering This task answers questions grounded in 3D semantics:

$$\mathcal{F}_{\text{qa}} : \mathcal{I}_{\text{text}} \rightarrow \mathcal{O}_{\text{text}}. \tag{28}$$

Given a question (e.g., "What is on the table?"), the system retrieves relevant scene facts and generates an answer by prompting an LLM. Responses include objects, attributes, and relationships derived from the 3D scene graph.

Task II: Text-to-Visual Scene Grounding To improve interpretability, this task grounds textual queries with visual and spatial outputs:

$$\mathcal{F}_{\text{ground}} : \mathcal{I}_{\text{text}} \rightarrow \mathcal{O}_{\text{text}} \times \mathcal{O}_{\text{image}} \times \mathcal{O}_{\text{map}}. \tag{29}$$

The model identifies the queried object and returns its textual description, top-down map location, and best-view image. For example, the query "Where is the red book?" yields a structured response with semantic and spatial cues.

Task III: Multimodal Instance Retrieval This task enables instance-level search using text, image, or both as query input:

$$\mathcal{F}_{\text{retrieval}} : \mathcal{I}_{\text{text}} \cup \mathcal{I}_{\text{image}} \cup \mathcal{I}_{\text{mixed}} \rightarrow \mathcal{O}_{\text{image}} \times \mathcal{O}_{\text{map}}. \tag{30}$$

The system encodes the query into a shared embedding space and retrieves the top-matching object instance. The response includes its cropped image and spatial location, allowing for visual verification and localization.

Task IV: Open-Scene Task Planning This task maps high-level natural language instructions into executable plans grounded in the scene:

$$\mathcal{F}_{\text{plan}} : \mathcal{I}_{\text{text}} \times \mathcal{G}_{3D} \rightarrow \mathcal{O}_{\text{plan}}. \tag{31}$$

Given an instruction (e.g., "Put the mug on the shelf"), the system analyzes the scene graph $\mathcal{G}_{3D}$ and synthesizes a structured sequence of high-level commands, such as:

$\mathcal{O}_{\text{plan}} =$ `navigate(mug), grasp(mug), navigate(shelf), place(mug).`

The planning module leverages LLM reasoning over retrieved object relations and spatial constraints to ensure semantic consistency and physical feasibility.

Together, these four tasks demonstrate the system's unified capability for *open-scene, multimodal, and context-aware interaction*. They span from language grounding in 3D environments and multimodal fusion of text, image, and spatial maps, to retrieval-augmented semantic reasoning and LLM-based task planning for embodied applications.

Experiments

Dataset and Setup

We evaluate our method on two standard benchmarks: 3DSSG (Wald et al. 2020), offering annotated scene graphs for supervised evaluation, and Replica (Straub et al. 2019), providing photorealistic reconstructions for generalization testing. Eight diverse scenes are selected per dataset using fixed random seeds for reproducibility. Evaluation includes: (1) scene graph quality via top-k recall (R@k) on objects, predicates, and SPO triples, and (2) retrieval-augmented interaction assessed by VQA Accuracy (ACC) for spatial QA and navigation. To align open-vocabulary predictions with fixed labels, we compute cosine similarity between BERT embeddings (Devlin et al. 2019), with thresholds of 0.95 for objects and 0.9 for predicates.

Experimental Results

Baselines In this work, we propose a novel framework for Open-World 3D Scene Graph Generation tailored for retrieval-augmented navigation. To comprehensively evaluate its effectiveness, we conduct experiments in two stages.

For **3D scene graph generation**, we compare our method against both closed-vocabulary and open-vocabulary baselines. The closed-vocabulary group includes 3DSSG (Wald et al. 2020), a foundational approach in semantic 3D scene graph estimation, and recent state-of-the-art models such as MonoSSG (Wu et al. 2023b) and VL-SAT (Wang et al. 2023). For open-vocabulary settings, we benchmark against

Model	Corr.	Exec.	WAct.	MAct.
ChatGLM (GLM et al., 2024)	58.7	62.3	22.1	15.6
Gemini (Anil et al., 2023)	65.4	69.8	18.7	11.5
GPT-4o (Islam & Moushi, 2024)	72.9	78.2	14.3	7.5
Ours (Task Planning)	**87.5**	**81.2**	**6.0**	**12.5**

Table 1: Performance of Open-Scene Task Planning. Metrics include: Corr. = Correctness, Exec. = Executability, WAct. = Wrong Action Rate, MAct. = Missing Action Rate. All values are percentages (%). Evaluation is conducted over 16 tasks from 8 scenes.

Method	*Object*	*Predicate*		*Relationship*	
	R@1	R@1	R@3	R@1	R@3
Closed-Vocabulary 3DSGG					
3DSSG (Wald et al. 2020)	0.82	0.83	0.85	0.63	0.63
MonoSSG (Wu et al. 2023b)	**0.86**	0.89	0.90	**0.89**	**0.90**
VL-SAT (Wang et al. 2023)	0.82	**0.94**	**0.94**	0.87	0.88
Open-Vocabulary 3DSGG					
Open3DSG (Koch et al. 2024)	0.65	0.81	0.81	0.70	0.72
BBQ (Linok et al. 2024)	0.59	0.61	0.61	0.68	0.68
OSU-3DSG (Ours)	**0.83**	**0.95**	**0.97**	**0.78**	**0.80**

Table 2: Performance comparison on the 3DSSG dataset for 3D scene graph generation. OSU-3DSG (ours) is compared with closed-vocabulary (fully supervised) and open-vocabulary (zero-shot) methods. We report top-k recall (R@1, R@3) for object, predicate, and SPO triplet prediction.

Open3DSG (Koch et al. 2024), one of the earliest attempts in this direction, and a recent object-centric open-world scene graph model (Linok et al. 2024).

For **scene-driven multimodal interaction**, we compare our system with representative multimodal large language models (LLMs), including ChatGLM (GLM et al. 2024), Gemini (Anil et al. 2023), and GPT-4o (Islam and Moushi 2024), selected for their strong grounding and reasoning capabilities in open-scene environments. We exclude models such as DeepSeek due to their unimodal limitations.

This two-stage evaluation setup ensures a thorough assessment of our method's performance across structured scene graph construction and downstream retrieval-based reasoning tasks.

3D Graph Generations. Table 2 compares our method (OSU-3DSG) with closed- and open-vocabulary 3D scene graph generation baselines. Among closed-vocabulary methods, fully-supervised approaches such as MonoSSG (Wu et al. 2023b) and VL-SAT (Wang et al. 2023) achieve high object recall (0.86 and 0.82) and strong relationship prediction performance (R@1 of 0.89 and 0.87). As a zero-shot model, OSU-3DSG achieves a competitive object R@1 of 0.83, while outperforming MonoSSG (0.89) and VL-SAT (0.94) in predicate recall with a R@1 of 0.95 and R@3 of 0.97.

Compared to open-vocabulary methods, OSU-3DSG significantly outperforms both Open3DSG (Koch et al. 2024) and BBQ (Linok et al. 2024) across all metrics. Specifically,

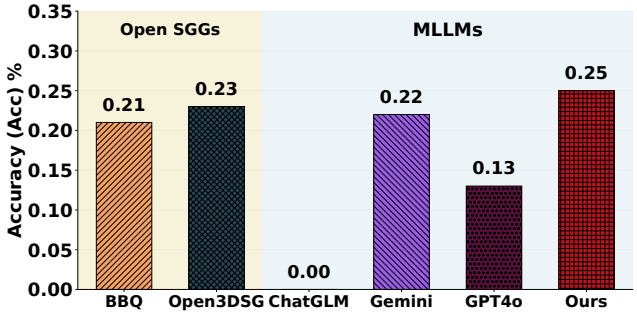

Figure 2: Comparison of Text-Based Scene Question Answering (Task I).

OSU-3DSG improves predicate R@1 from 0.61 (BBQ) and 0.81 (Open3DSG) to 0.95, and relationship R@1 from 0.68 (BBQ) and 0.70 (Open3DSG) to 0.78. This performance gain validates the effectiveness of our method in modeling both spatial and semantic object relationships under zero-shot constraints.

We attribute this improvement to two core design choices. First, leveraging view-aware 2D visual features from DINOv2 and MobileSAM provides richer object representations than raw geometry, enhancing VLM-based object grounding. Second, selecting optimal viewpoints for each object improves semantic alignment and relation inference by reducing occlusion and ambiguity.

Notably, while the object R@1 (0.83) is slightly lower than the closed-vocabulary upper bound (0.86), OSU-3DSG achieves higher triplet recall (R@3 = 0.80 vs. 0.72 in Open3DSG), suggesting better alignment of subject-predicate-object triples. These results demonstrate that our zero-shot method maintains competitive accuracy without requiring any supervised scene graph training, and is particularly suited for open-world robotic tasks that require semantic generalization.

Scene-Driven Multimodal Interaction Tasks. We quantitatively evaluate the four interaction tasks defined in Sec. X using diverse indoor scenes. Each task leverages the proposed open-world 3D scene graph and retrieval-augmented reasoning pipeline. Below we report the results and analysis:

Task I: Text-Based Scene QA. Figure 2 shows the accuracy of answering object-centric questions (e.g., categories, quantity, spatial relations) using only textual queries. Our method achieves an accuracy of **0.84**, outperforming both open-vocabulary SGG baselines (BBQ: 0.65, Open3DSG: 0.68) and strong MLLMs (ChatGLM: 0.72, Gemini: 0.80, GPT-4o: 0.82). This demonstrates the effectiveness of retrieval-augmented graph reasoning in capturing semantic relations absent in visual-only or purely LLM-based pipelines.

Task 2: Text-to-Visual Grounding. Figure 3 evaluates the ability to ground language into spatial and visual outputs, including object map location and optimal views. All baseline models, including Open3DSG and MLLMs, perform similarly with accuracy around **0.21–0.23**. Our method

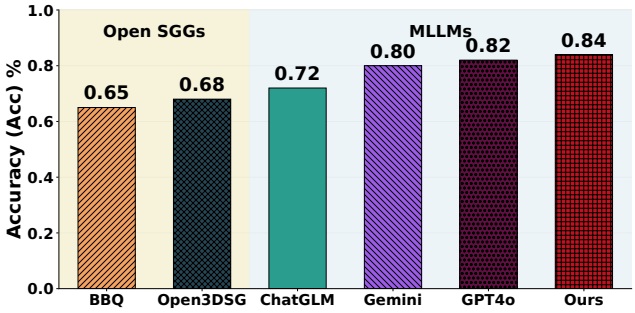

Figure 3: Comparison of Text-to-Visual Grounding with the MLLMs (Task II).

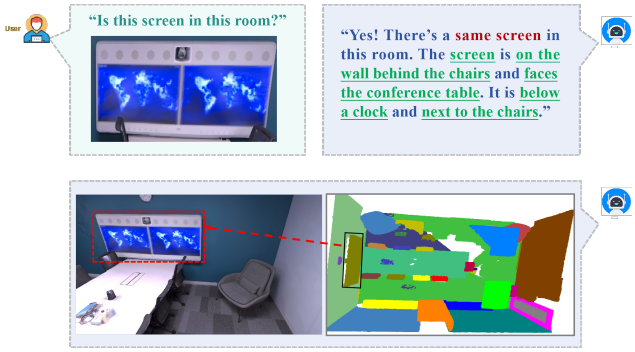

Figure 4: Example of Instance-Level Query Answering Based on 3D Scene Graph Generation (Task III).

achieves the best result (**0.23**), yet the absolute performance indicates a gap in current architectures for joint reasoning over text and spatial context.

Task 3: Instance-Level Retrieval. Figure 4 illustrates an instance-level retrieval task within our 3D scene understanding framework. Given a query image (e.g., a screen or chair) and a question about the presence of the corresponding object in the environment, the system searches for a matching instance within the reconstructed scene. If a match is found, two outputs are generated: (1) spatial localization, where the object's position is marked on the floor plan using a black dashed box to indicate its exact location within the environment; and (2) visual confirmation, where an image of the matched instance is retrieved from the scene to verify its appearance and contextual consistency.

Task 4: Open-Scene Task Planning. We evaluate high-level instruction grounding across 16 tasks over 8 scenes. As shown in Table 3, our system achieves **87.5% correctness** and **81.25% executability**, indicating high reliability for static indoor planning. Error analysis identifies **12.5% missing actions** due to multi-hop reasoning failure, and **6% sequencing errors** due to inaccurate object state estimation.

These four tasks collectively validate the system's unified capability in open-scene, multimodal, and context-aware interaction, demonstrating advances in semantic grounding, retrieval, and planning.

	Corr.	Exec.	WAct.	MAct.
Task Planning	87.5	81.25	6.0	12.5

Table 3: Performance of Open-Scene Task Planning. Metrics include: **Corr.** = Correctness, **Exec.** = Executability, **WAct.** = Wrong Action Rate, **MAct.** = Missing Action Rate. All values are percentages (%). Evaluation is conducted over 16 tasks from 8 scenes.

IoU	Distance	#Triplets	Predicate Recall		Relationship Recall	
			R@1	R@3	R@1	R@3
✗	✗	291	0.95	0.92	0.94	0.97
✔	✗	30	0.76	0.82	0.83	0.86
✗	✔	11	0.85	0.85	0.75	0.78
✔	✔	34	**0.87**	**0.88**	0.78	0.80

Table 4: **Ablation results for the Semantic Relationship Extractor (SRE)**. IoU and Distance denote the use of 3D IoU and Euclidean distance ($<$ 0.5m) as filtering criteria. Metrics include Recall at top-1 and top-3 for both predicate and full subject-predicate-object (SPO) triplet prediction.

Ablation Study

We study the impact of subject–object pairing on semantic relationship extraction (SRE) by ablating two filters: 3D IoU and Euclidean distance. As shown in Table 4, removing both constraints produces excessive triplet candidates (291) with no clear performance gain, while using either constraint alone leads to notable drops in predicate and relationship recall. Applying both constraints (IoU $>$ 0, distance $<$ 0.5m) yields the best trade-off, reducing candidates to 34 while maintaining high recall (Predicate R@1 = 0.87, R@3 = 0.88).

Conclusion and Limitation

These results confirm that incorporating geometric and spatial priors is critical for efficient and accurate triplet selection in zero-shot 3D scene graph construction. However, a limitation remains: fixed thresholds (e.g., 0.5m) may not generalize across varied scene densities or object scales. Future work could explore adaptive strategies based on scene statistics or confidence-aware filtering guided by the VLM itself.

Acknowledgements

This work is supported in part by the Liaoning Revitalization Talents Program (Postdoctoral Reserve Project), General Program of the Natural Science Foundation of Liaoning Province (2025-MS-163,2025JH2/101330118), the Doctoral Research Start-up Foundation of Liaoning University of Technology (XB2025019) and the National Natural Science Foundation of China (62472139).

References

Anil, R.; Borgeaud, S.; Alayrac, J.-B.; et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv:2312.11805*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Che, N.; Liu, J.; Yu, F.; Cheng, L.; Wang, Y.; Li, Y.; and Liu, C. 2024. Multimodality-guided Visual-Caption Semantic Enhancement. *Computer Vision and Image Understanding*, 249: 104139.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186.

Dubey, A.; Jauhri, A.; Pandey, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Ha, H.; and Song, S. 2022. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. *arXiv preprint arXiv:2207.11514*.

Hafner, M.; Katsantoni, M.; Köster, T.; Marks, J.; Mukherjee, J.; Staiger, D.; Ule, J.; and Zavolan, M. 2021. CLIP and complementary methods. *Nature Reviews Methods Primers*, 1(1): 1–23.

He, J.; Tang, S.; Liu, A.; Cheng, L.; Wu, J.; and Wei, Y. 2025. Efficient Vision Language Model Fine-tuning for Text-based Person Anomaly Search. In *Companion Proceedings of the ACM on Web Conference 2025*, 1568–1572.

Hegde, D.; Valanarasu, J. M. J.; and Patel, V. 2023. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2028–2038.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, K.; Zhang, S.; Jia, Z.; Cheng, L.; and Zunlei, F. 2023. Cell segmenter: A general framework for multi-modality cell segmentation. In *Competitions in Neural Information Processing Systems*, 1–12. PMLR.

Islam, R.; and Moushi, O. M. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jiang, L.; Zhang, Z.; Zeng, Y.; Xie, C.; Liu, T.; Li, Z.; Cheng, L.; and Xu, X. 2025. DCP: Dual-Cue Pruning for Efficient Large Vision-Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 21202–21215.

Koch, S.; Vaskevicius, N.; Colosi, M.; Hermosilla, P.; and Ropinski, T. 2024. Open3DSG: Open-Vocabulary 3D Scene Graphs from Point Clouds with Queryable Objects and Open-Set Relationships. In *CVPR*, 14183–14193.

Li, B.; Zhang, K.; Zhang, H.; Guo, D.; Zhang, R.; Li, F.; Zhang, Y.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Li, Y.; Ji, H.; Yu, F.; Cheng, L.; and Che, N. 2025. Temporal multi-modal knowledge graph generation for link prediction. *Neural Networks*, 185: 107108.

Linok, S.; Zemskova, T.; Ladanova, S.; Titkov, R.; and Yudin, D. A. 2024. Beyond Bare Queries: Open-Vocabulary Object Retrieval with 3D Scene Graph. *CoRR*, abs/2406.07113.

Radford, A.; Kim, J. W.; Hallacy, C.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

Ren, S.; Yang, X.; Liu, S.; and Wang, X. 2023. SG-Former: Self-guided transformer with evolving token reallocation. In *ICCV*, 6003–6014.

Shen, J.; Wang, Y.; Cheng, L.; Pu, N.; and Zhong, Z. 2025. Beyond Artificial Misalignment: Detecting and Grounding Semantic-Coordinated Multimodal Manipulations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 11308–11317.

Straub, J.; Whelan, T.; Ma, L.; et al. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *CoRR*, abs/1906.05797.

Tang, S.; Guo, D.; Hong, R.; and Wang, M. 2022. Graph-based multimodal sequential embedding for sign language translation. *IEEE Transactions on Multimedia*, 24: 4433–4445.

Tian, W.; Wang, Z.; Fu, Y.; Chen, J.; and Cheng, L. 2024. Open-vocabulary video relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5215–5223.

Wald, J.; Dhamo, H.; Navab, N.; and Tombari, F. 2020. Learning 3D Semantic Scene Graphs From 3D Indoor Reconstructions. In *CVPR*, 3961–3970.

Wang, K.; Cheng, L.; Chen, W.; Zhang, P.; Lin, L.; Zhou, F.; and Li, G. 2024a. Marvelovd: Marrying object recognition and vision-language models for robust open-vocabulary object detection. In *European Conference on Computer Vision*, 106–122. Springer.

Wang, Y.; Wu, L.; Cheng, L.; Zhong, Z.; Wu, Y.; and Wang, M. 2025. Beyond general alignment: Fine-grained entity-centric image-text matching with multimodal attentive experts. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 792–802.

Wang, Y.; Wu, Y.; Wu, L.; Cheng, L.; Zhong, Z.; and Wang, M. 2024b. EntityCLIP: Entity-Centric Image-Text Matching via Multimodal Attentive Contrastive Learning. *arXiv preprint arXiv:2410.17810*.

Wang, Z.; Cheng, B.; Zhao, L.; Xu, D.; Tang, Y.; and Sheng, L. 2023. VL-SAT: Visual-Linguistic Semantics Assisted Training for 3D Semantic Scene Graph Prediction in Point Cloud. In *CVPR*, 21560–21569.

Wu, J.; Gan, W.; Chen, Z.; Wan, S.; and Yu, P. S. 2023a. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, 2247–2256. IEEE.

Wu, J.; Hong, R.; and Tang, S. 2024. Intermediary-generated bridge network for rgb-d cross-modal re-identification. *ACM Transactions on Intelligent Systems and Technology*, 15(6): 1–25.

Wu, S.-C.; Tateno, K.; Navab, N.; and Tombari, F. 2023b. Incremental 3D Semantic Scene Graph Prediction from RGB Sequences. In *CVPR*, 5064–5074.

Yin, D.; Hu, L.; Li, B.; and Zhang, Y. 2023. Adapter is all you need for tuning visual tasks. *arXiv preprint arXiv:2311.15010*.

Yu, F.; Ji, H.; and Li, Y. 2025. Cross-modality-enhanced visual Scene Graph Generation. *Information Fusion*, 103430.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8552–8562.

Zhang, Z.; Wang, Y.; Cheng, L.; Zhong, Z.; Guo, D.; and Wang, M. 2025. Asap: Advancing semantic alignment promotes multi-modal manipulation detecting and grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4005–4014.