OP3DSG: Open-Vocabulary Part-Aware 3D Scene Graph Generation for Real-World Environments

Yirum Kim[1] and Ue-Hwan Kim[1⋆]

Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea
kimyirum@gm.gist.ac.kr, uehwan@gist.ac.kr

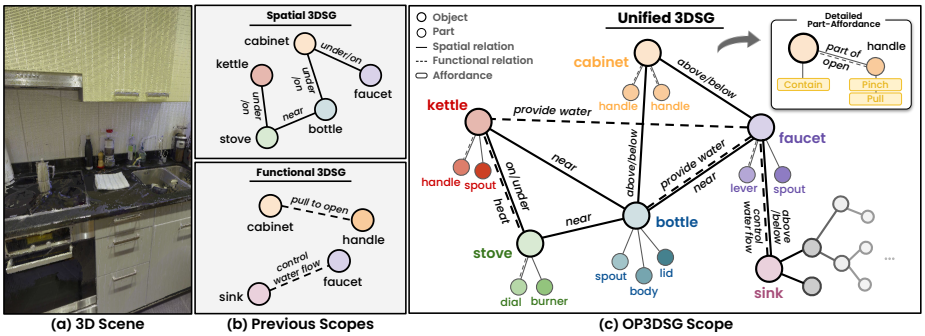

Fig. 1: Conceptual Comparison of 3D Scene Graphs (3DSGs). (a) Given a 3D scene, (b) prior work falls into two scopes: *Spatial 3DSGs*, which recognize objects and encode spatial relations, and *Functional 3DSGs*, which mainly recognize interactive parts and functional relations. (c) Our scope unifies these notions into *Unified 3DSGs*.

Abstract. 3D scene graphs (3DSGs) provide a compact and structured abstraction of 3D environments. Although advances in foundation models have enabled open-vocabulary 3DSG generation, existing approaches remain object-centric and encode limited relational information—restricting their applicability in real-world scenarios that require fine-grained understanding. We propose OP3DSG, an open-vocabulary part-aware 3DSG generation framework that constructs unified graphs that jointly model objects, interactive parts, spatial relations, functional relations, and affordances. OP3DSG integrates object-part knowledge-guided detection with part-aware 3D fusion to preserve small and interaction-relevant components, and employs a geometry-initialized prior graph with LLM-based refinement to reduce spurious relational predictions while enabling efficient graph construction. To systematically evaluate unified 3D scene graph construction, we introduce UniGraph3D, a benchmark designed for part-aware perception and multi-level relational reasoning. Experimental results show that OP3DSG achieves state-of-the-art performance and demonstrates its effectiveness as a perception backbone in diverse real-world robotics tasks.

Keywords: 3D Scene Graph · Embodied AI · Scene Understanding

⋆ Corresponding author.
Project page: https://k2room.github.io/OP3DSG

1 Introduction

3D scene graphs (3DSGs) are a compact, structured abstraction that connects geometry with semantics and relationships, enabling efficient querying and reasoning over large environments [19, 46, 54, 55]. Such structure is particularly attractive for embodied agents—conducting various downstream tasks such as navigation [12,23] and task planning [2,35]—where a graph can serve as a persistent, queryable "mental model" over long horizons and across viewpoints [2, 12]. Recent progress has pushed 3DSGs toward *open-vocabulary* settings by leveraging vision–language foundation models and open-set detectors [21, 27, 63]. This enables graphs to be queried with novel category names beyond closed label spaces—leading to open-vocabulary 3DSG construction pipelines [14,20]. In parallel, a complementary line of work argues that *functionality* must be part of the representation—proposing functional 3DSGs that model interactive elements and functional relations [8, 59].

Despite these advances, current open-vocabulary 3DSGs remain largely *object-centric* and often fail where physical interaction demands precision: small components (*e.g.*, buttons, handles, switches) and fine-grained dependencies (*e.g.*, a remote controlling a TV, a knob operating a burner) are missing or poorly grounded. On the other hand, the function-centric representation remains incomplete for general-purpose embodied reasoning: spatial structure, object-part hierarchy, and affordance are often handled in separate modules. Consequently, task planning that reasons over object-level graphs can navigate to a "microwave" but cannot decompose "heat food" into sub-goals like "press start button," whereas reasoning over function-centric graphs does not generate high-level plans. Put simply, embodied agents require a *compositional* view of the world: not only *what* is present, but *which part* is actionable and *how* that part enables a function.

We take a step toward this goal by introducing the **Unified 3D Scene Graph** (Fig. 1): a single graph that jointly models *objects*, *interactive parts*, *functional relations*, *spatial relations*, and *affordances*. Constructing such unified 3DSGs, however, is fundamentally challenging for two reasons. First, *part-level open-vocabulary grounding is intrinsically harder than object-level grounding*: parts are smaller, more ambiguous, heavily viewpoint-dependent, and systematically under-represented in the predominant object-centric data distributions of modern foundation models [21, 25, 27, 37, 61, 63]—often producing object-level localization even with part-centric prompts

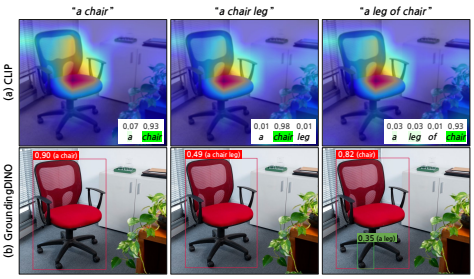

Fig. 2: Preliminary experiments reveal limitations in part detection. (a) With CLIP [33], part-centric phrases (*e.g.*, *"a chair leg", "a leg of chair"*) yield activation maps nearly identical to the object-level prompt *"a chair"*, as attention concentrates on the token *chair*. (b) With GroundingDINO [27], part-centric phrases still yield object-level bounding boxes; when a part is detected, its confidence score is notably low.

(Fig. 2). Further, transferring robust part grounding into unconstrained *3D* reconstruction pipelines remains brittle even with dedicated part datasets [16,34]. Second, unification *explodes the reasoning space*: once parts become nodes, naive relation prediction scales quadratically with the number of entities, quickly becoming computationally expensive and increasingly error-prone if one relies on unconstrained LLM inference over all candidate pairs—leading to spurious functional relations or missed dependencies.

To address these challenges, we propose ***OP3DSG***, the **O**pen-vocabulary **P**art-aware **3D S**cene **G**raph generation framework for real-world environments. OP3DSG makes unified graphs feasible by combining: (i) *knowledge-guided control of the entity space* that expands and validates part candidates via object-part structure, (ii) *part-aware multi-view 3D fusion* that preserves fine-grained geometry for small components across viewpoints, and (iii) *geometry-anchored, verification-gated LLM reasoning* that scales relation/affordance inference without exhaustive node/edge enumeration while avoiding hallucinations. Our key insight is that *geometry can act as a strong prior that makes open-vocabulary, part-aware reasoning tractable*. Instead of inferring a unified graph from scratch, we first build a geometry-initialized prior graph that captures stable spatial neighborhood structure and object-part hierarchy from multi-view fusion. We then use an LLM not as a free-form generator, but as a *verification-gated refiner*: it proposes and validates only those relations and affordances that are plausible under geometric constraints, turning unstructured language reasoning into constrained refinement.

To systematically evaluate the unified 3D scene graph generation task, we build the *UniGraph3D* benchmark. UniGraph3D consolidates the function-centric 3DSG datasets [8,59] and augments them with annotations for interactive object-part structure, spatial and functional relations, and affordances. Our benchmark not only supports fair comparison across prior methods but also facilitates future work on fine-grained 3D scene understanding. Experiments show that our OP3DSG substantially improves part-level grounding and multi-level relation prediction in indoor environments, and that unified graphs serve as a strong perception backbone for downstream embodied tasks such as language-conditioned querying [14, 43], navigation [23], and task planning [2, 35].

In summary, we make three contributions:

- **Problem Formulation.** We introduce the *Unified 3D Scene Graph* generation task, which jointly models objects, interactive parts, spatial and functional relations, and affordances within a single representation, enabling fine-grained embodied reasoning.
- **Model Design.** We propose *OP3DSG*, the unified 3D scene graph generation framework integrating knowledge-guided part grounding, fine-grained 3D fusion, and geometry-anchored, verification-gated reasoning to mitigate hallucinations.
- **Benchmark Setup.** We introduce *UniGraph3D*, a benchmark with consolidated datasets, enriched annotations, and evaluation protocols for systematic assessment of unified 3D scene graph generation.

2 Related Work

3D Scene Graph. 3DSGs that encode semantic information can be categorized into two research directions: *Spatial 3DSGs* and *Functional 3DSGs*.

(*i*) *Spatial 3DSGs* represent objects as nodes and model spatial relationships between objects as edges. Early works [3, 11, 19, 46, 54, 55] adopt this paradigm, constructing graphs based on geometric consistency and object detection from point clouds or RGB-D inputs. To support higher-level abstraction, hierarchical 3DSG methods [3, 5, 18, 38] organize scenes into multi-level graphs (*e.g.*, building–room–object hierarchies), enabling scalable reasoning over large environments. With the emergence of 2D foundation models [21, 27, 33, 37], recent approaches [4, 14, 20, 52] have extended spatial 3DSG generation to an open-vocabulary setting. Open3DSG [20] co-embeds geometric features with VLM embeddings, enabling zero-shot prediction of both object and relations. ConceptGraphs [14] proposes the training-free pipeline by leveraging 2D foundation models and multi-view association. Despite this progress, existing methods remain largely object-centric and tend to struggle with small or thin structures, leading to incomplete recognition of fine-grained parts. As a result, spatial 3DSGs may overlook interactive parts that are critical in embodied scenarios [2, 23, 35].

(*ii*) *Functional 3DSGs* focus on modeling functional relations between objects or parts. As an emerging line of research, recent approaches [39, 59] leverage VLMs and LLMs to encode functional knowledge by generating rich language descriptions for each node to capture interaction semantics. However, existing methods still exhibit limited part-level perception ability and do not encode spatial structure. Moreover, relying on VLM-based captioning to compensate for insufficient detection granularity introduces substantial computational overhead and prolongs the graph construction time.

Considering spatial and functional graphs in isolation yields incomplete scene representations. We therefore move toward Unified 3DSGs, providing a comprehensive representation that jointly models part-level nodes with spatial and functional relations.

Part-level Perception. Beyond object-level recognition [21, 31, 56], part-level segmentation [22, 28, 49] aims to achieve fine-grained understanding by recognizing the constituent components of objects. With the emergence of large-scale benchmarks [16, 30, 32, 34, 62] for common objects, research on part-level perception has expanded to more diverse object categories [32, 41, 51]. Subsequent works have further extended part-level understanding into the 3D domain, either by lifting 2D predictions into 3D space [13,26,29,57,64] or by directly performing part segmentation on detailed point cloud representations [47, 48, 58]. Building upon recent advances in part-level perception and 2D–3D lifting techniques, we extend part-level understanding of common indoor objects to the 3D domain. Specifically, we project part-level predictions into 3D space and associate them with geometric structures to ensure spatial consistency. Based on these part-aware representations, we construct part-aware 3D scene graphs, enabling structured and fine-grained scene understanding.

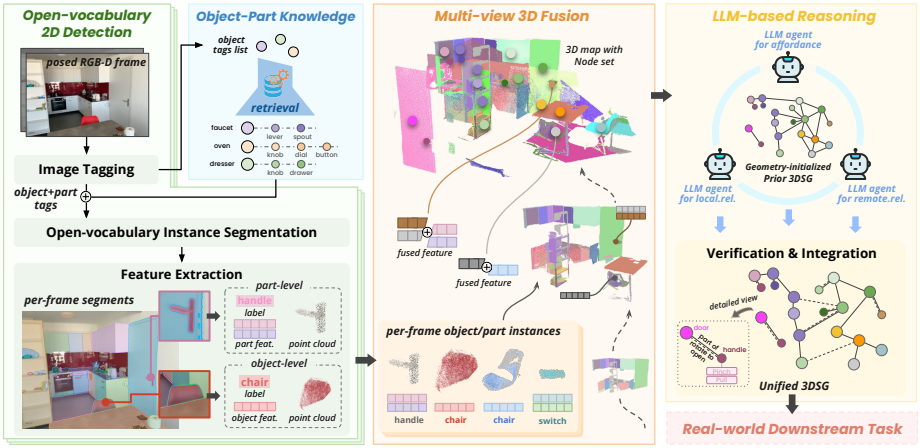

Fig. 3: Overview of the proposed framework. Knowledge-guided 2D detection generates object and part instances that are incrementally fused into a global 3D map. The resulting nodes define a geometry-initialized prior graph, which is refined by LLM-based reasoning into the Unified 3DSG.

3 Problem Formulation

Given a posed RGB-D sequence $\{I_t, D_t\}_{t=0}^{T}$, where I_t denotes the RGB frame and D_t the corresponding depth, our goal is to construct a unified 3D scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that encodes objects, parts, spatial relations, functional relations, and affordances. The vertex set $\mathcal{V} = \mathcal{V}^o \cup \mathcal{V}^p$ consists of two types of nodes: object nodes $\mathcal{V}^o$, representing physical objects in the environment, and interactive part nodes $\mathcal{V}^p$, representing sub-components of objects (*e.g.*, *button*, *handle*, *seat*). Each node $v \in \mathcal{V}$ has a set of affordance labels $\mathcal{A}^v$ (*e.g.*, *rotate*, *pinch pull*, *tip push*), which specifies feasible physical interactions. The edge set $\mathcal{E} = \mathcal{E}^s \cup \mathcal{E}^f$ encodes both spatial relations $\mathcal{E}^s$ (*e.g.*, *on*, *next to*, *part of*) and functional relations $\mathcal{E}^f$ (*e.g.*, *turn on/off*, *control flow*, *remote*).

4 Method

4.1 Overview

We propose OP3DSG, a model designed for part-level perception and unified 3D scene graph generation in an open-vocabulary manner. As illustrated in Fig. 3, the framework consists of three main components: *(i) object-part 2D detection*, *(ii) multi-view 3D fusion*, and *(iii) LLM-based reasoning*.

In the object/part detection stage (Sec. 4.2), we leverage a foundation model pretrained on part-centric datasets to handle the challenge of fine-grained part recognition. By incorporating object–part knowledge, the model ensures comprehensive perception coverage and mitigates the omission of small or functionally important parts. The 3D fusion stage (Sec. 4.3) incrementally integrates

the segmentation results from each frame into a global 3D map. To accurately merge small parts observed from multiple views, we propose a fine-grained fusion strategy that jointly considers geometric, chromatic, and semantic consistency. Finally, in the LLM-based reasoning stage (Sec. 4.4), multiple LLM agents take the geometry-initialized prior 3DSG and each textual prompt as input. Unlike VLM-based methods, whose efficiency degrades as the number of objects grows, we adopt a language-only reasoning architecture with the geometry-anchored verification gate to avoid scalability bottlenecks while maintaining robust relational reasoning.

4.2 Object-Part 2D Detection

Object/Part Candidates. Despite the success of foundation models in object detection, small objects or parts remain challenging. To address this gap, we introduce a simple strategy driven by object–part knowledge-guided control of the entity space. Given an RGB frame F_t, an open-vocabulary image tagging model [60] yields initial object candidates C_t^o. A pre-built object–part knowledge base $\mathcal{K}$, empirically curated to cover common object–part relations in indoor scenes, is used to control the entity space. For each detected object label $c \in C_t^o$, we retrieve a predefined set of plausible interactive parts $\mathcal{K}(c)$ (*e.g.*, *oven* $\rightarrow$ {*oven button*, *oven door*, *oven handle*}), forming the part candidate set $C_t^p = \bigcup_c \mathcal{K}(c)$. The final prompt list is constructed as $C_t = C_t^o \cup C_t^p$. This knowledge base guides part-level prompting without restricting the open-vocabulary nature of object recognition—for example, if *door* tag is generated but its *handle* or *knob* is not recognized, we explicitly include the corresponding part tags before performing detection.

Furthermore, we adopt the open-vocabulary detection model [41] pre-trained on large-scale part-centric datasets [15,34]. Prompted with C_t, the detector outputs semantic label $\{l_{t,i}\}_{i=1}^N$ and masks $\{m_{t,i}\}_{i=1}^N$ for the N detected instances. Each mask is then back-projected to 3D and transformed into the global map, yielding the per-instance point cloud $\mathrm{p}_{t,n}$. Overall, the knowledge base provides a simple yet effective way to inject part-level knowledge to compensate for insufficient part recognition, while preserving the open-vocabulary nature of the detection modules.

Feature Extraction. For object association in 3D fusion, we extract region-level semantic features $\{f_{t,i}^s\}_{i=1}^N$ from the recognition head of the detector, which align with the CLIP [33] text embedding space. However, relying solely on semantic information is often inadequate for part-level association. Unlike objects that may comprise multiple heterogeneous materials, parts are typically composed of a consistent material. Motivated by this observation, we hypothesize that color and texture cues can provide more stable evidence for part association. Hence, we employ classical approaches [7, 24, 42, 45] to compute color-distribution features $f_{t,i}^c$ for part instances. For each part mask $m_{t,i}^p$, we apply white balancing to reduce illumination bias and define a valid pixel set $\Omega_{t,i}$ by excluding specular

highlights. Specifically, we discard pixels that are simultaneously over-exposed and weakly chromatic, while keeping the remaining masked pixels:

$$\Omega_{t,i} = \{(x,y) \mid m^p_{t,i}(x,y) = 1, (S(x,y) > \tau_S) \wedge (V(x,y) < \tau_V)\}\,, \tag{1}$$

where S and V denote the saturation and value components in the HSV color space, respectively. The saturation threshold τ_S and the intensity threshold τ_V jointly define a specular-like outlier region. The color-distribution feature $f^c_{t,i}$ is concatenated with three normalized histogram vectors representing complementary color spaces:

$$f^c_{t,i} = [\, H^{\mathrm{cn}}(\Omega_{t,i}),\, H^{\mathrm{ab}}(\Omega_{t,i}),\, H^{\mathrm{opp}}(\Omega_{t,i})\,]\,, \tag{2}$$

where H^{cn} is based on Color Names [7,45], H^{opp} on the Opponent color space [10, 44], and H^{ab} on the CIELAB a^*b^* space.

4.3 Multi-view 3D Fusion

Object Association. For every newly-detected object $v^o_{t,i} = (f^s_{t,i}, \mathrm{p}_{t,i})$, we compute both geometric and semantic similarities against the existing objects in the map $\{v^o_{t-1,j}\}_{j=1}^{J}$, where $v^o_{t-1,j} = (f^s_{t-1,j}, \mathrm{p}_{t-1,j})$ might partially overlap in geometry. We define geometric similarity s^g as the nearest neighbor ratio (NNR), while semantic similarity s^s is the cosine similarity between L2-normalized descriptors and is affine-transformed to $[0,1]$ [14]. The aggregated similarity S^o between the object i and j is defined as

$$s^g(i,j) = NNR(\mathrm{p}_{t,i}, \mathrm{p}_{t,j}),\ s^s(i,j) = (f^{s\top}_{t,i} f^s_{t,j} + 1)/2, \tag{3}$$

$$S^o(i,j) = s^g(i,j) + s^s(i,j)\,. \tag{4}$$

We adopt a greedy online association strategy for incremental object fusion. For each newly detected object instance, we associate it with the existing objects that achieve the highest similarity. If the highest similarity is below a predefined threshold, a new object node is initialized.

Part Association. For every newly-detected part $v^p_{t,i} = (f^s_{t,i}, f^c_{t,i}, \mathrm{p}_{t,i})$, we additionally compute color-distribution similarities against the existing parts in the map $\{v^p_{t-1,j}\}_{j=1}^{J}$, where $v^p_{t-1,j} = (f^s_{t-1,j}, f^c_{t-1,j}, \mathrm{p}_{t-1,j})$ that partially overlap in geometry. The color-distribution similarity s^c between parts i and j is defined as a weighted sum of the chi-square distances d^c computed on each color histogram:

$$d^c(i,j) = w^c_1 \cdot d^{\mathrm{cn}}(i,j) + w^c_2 \cdot d^{ab}(i,j) + w^c_3 \cdot d^{\mathrm{opp}}(i,j)\ ,\ s^c(i,j) = \frac{1}{1 + d^c(i,j)}, \tag{5}$$

where $d^{\mathrm{cn}}, d^{ab}, d^{\mathrm{opp}}$ denote the χ^2 distances between the corresponding histogram $H^{\mathrm{cn}}, H^{\mathrm{ab}}, H^{\mathrm{opp}}$, respectively, and w^c_k denotes the weight assigned to each component. The aggregated similarity S^p between part i and j is defined as

$$S^p(i,j) = w^p \cdot \big(s^g(i,j) + s^s(i,j)\big) + (1 - w^p) \cdot s^c(i,j)\,, \tag{6}$$

where w^p denotes the weights for geometric/semantic and color-distribution similarity terms. We follow the same greedy assignment of object association to enable efficient streaming fusion without requiring a global optimal assignment.

Feature Update. For each associated node $v^o_{t-1,j}$ and $v^p_{t-1,j}$, the corresponding features are updated incrementally with a new observation. We update the semantic feature by averaging the previous and current representations, while the geometric information by taking the union of point clouds by down-sampling to remove duplicate points [14]:

$$f^s_{t,j} \leftarrow \frac{n_{t-1,j} \cdot f^s_{t-1,j} + f^s_{t,i}}{n_{t-1,j} + 1}\ , \quad \mathrm{p}_{t,j} \leftarrow \mathrm{p}_{t-1,j} \cup \mathrm{p}_{t,i}\ , \tag{7}$$

where $n_{t-1,j}$ denotes the number of detections previously associated with $v^o_{t-1,j}$ or $v^p_{t-1,j}$. For the color-distribution feature, each histogram is updated according to the exponential moving average (EMA) rule. For each histogram $H^{(k)}$ with $k \in \{\mathrm{cn}, \mathrm{ab}, \mathrm{opp}\}$, the update is defined as

$$\tilde{H}^{(k)}_t = (1-\alpha)\, H^{(k)}_{t-1} + \alpha\, \widehat{H}^{(k)}_t, \quad H^{(k)}_t = \frac{\tilde{H}^{(k)}_t}{\|\tilde{H}^{(k)}_t\|_1 + \varepsilon}\ , \tag{8}$$

where $\alpha \in (0,1)$ is a smoothing factor, $\|\cdot\|_1$ denotes the L^1 norm, and $\varepsilon > 0$ is a small constant for numerical stability.

4.4 LLM Reasoning

Prior 3D Scene Graph Generation. To compensate for the limited spatial perception capability of LLMs, we first construct a prior graph instead of directly encoding raw 3D coordinate information of nodes into the input prompt. From the final reconstructed 3D map $\{v_{T,j}\}_{j=1}^{J}$, each node $v_{T,j} = (l_{T,j}, \mathrm{p}_{T,j})$ consists of a semantic label $l_{T,j}$ and a point cloud $\mathrm{p}_{T,j}$. The label $l_{T,j}$ is determined by a voting scheme that selects the most frequent detection label, while $\mathrm{p}_{T,j}$ is used to compute the 3D center coordinates and bounding box size of the node.

To construct edges, we treat nodes whose 3D centers are within 1 meter as neighbors, *i.e.*, edge candidates. For inter-object pairs, we assign spatial relations (*e.g.*, *on*, *next to*) based on their relative geometric configurations. To model intra-object connectivity, we further apply semantic proximity matching: for part nodes spatially-close to multiple object nodes, we retain only the edge whose connected object label is semantically consistent with the part label—*i.e.*, when the object label text is contained within the part label text—and prune the remaining connections. Such a lexical heuristic is well-aligned with the knowledge-guided control, which generates part candidates in an object-conditioned form (*e.g.*, *cabinet handle*)—promoting coherent object–part hierarchies while suppressing ambiguous attachments.

Geometry-anchored LLM-based Reasoning. Given the prior 3DSG, we employ multiple LLM agents in parallel for reasoning. This multi-agent design decomposes the reasoning space—mitigating spurious predictions and explicitly separating short-range structural reasoning from long-range semantic reasoning. Each agent follows a step-wise prompt inspired by Chain-of-Thought [50]. Before each reasoning, all agents *independently* refine spatial relations in the prior graph: they validate inter-/intra-object relations by checking consistency with node labels, bounding-box extents, and 3D center coordinates derived from the prior 3DSG. This geometry-anchored verification filters spatially implausible relations and yields a physically consistent context for subsequent reasoning, without exhaustively enumerating all node pairs. The three agents are:

(i) Local relation reasoning agent infers local functional relations, which typically arise between physically connected object–part pairs (*e.g.*, *knob – pull to open – cabinet*). Rather than iterating over all $O(N^2)$ node pairs, the agent conditions on the prior 3DSG and reasons over the spatially associated object–part pairs in a single prompt, ensuring scalability as the number of nodes increases.

(ii) Remote relation reasoning agent handles functionally related but spatially distant entity pairs (*e.g.*, *TV – turn on/off – remote control*) by leveraging the LLM's pretrained world knowledge. To reduce the LLM's burden, the agent selects a set of remotable candidates from the prior graph and then predicts relation labels among these candidates, instead of considering all possible pairs.

(iii) Affordance reasoning agent predicts feasible physical interactions for each node (*e.g.*, *handle – pinch pull*, *bottle – contain*). For part nodes, affordance inference additionally conditions on the connected object label and geometric attributes, enabling context-aware and physically grounded predictions.

5 Experiment

5.1 UniGraph3D

To evaluate unified 3D scene graph generation, we propose *UniGraph3D* by consolidating and extending the FunGraph3D [59] and SceneFun3D† [8,59] datasets under a unified annotation schema. Rather than merely merging existing resources, our human annotators carefully re-examine the annotations and identify substantial labeling inconsistencies and ambiguities (*e.g.*, *"rotate to adjust the setting"* vs. *"rotate or press to adjust the setting"*). To address these issues, we consolidate ambiguous labels into a single canonical form, thereby enabling fair and consistent comparison across methods.

Beyond annotation refinement, UniGraph3D significantly expands the original datasets by introducing explicit object–part spatial relations and comprehensive affordance annotations for each node. Human annotators construct object–part connectivity based on the object and part labels with their coordinates. The affordances are assigned by aligning with existing work [8] while considering the geometric characteristics of each node, resulting in nine categories. As a result, UniGraph3D supports systematic assessment of part-aware perception and

Table 1: Comparison of 3DSG evaluation datasets. We compare the number of objects, parts, spatial/functional relations, and affordances. The sign † denotes the post-processed dataset [8] by additionally annotated functional relation [59].

Dataset	Object Node	Part Node	Spat. Edge	Func. Edge	Affordance
ReplicaSSG [17]	786	0	309	0	0
FunGraph3D [59]	146	201	0	228	0
SceneFun3D† [8, 59]	105	212	0	195	212
UniGraph3D (Ours)	251	414	357	425	554

multi-level relational reasoning, which were previously evaluated in isolation. As reported in Tab. 1, UniGraph3D provides a coherent and reliable testbed for unified 3D scene graph generation. Detailed statistics and procedures are provided in the supplementary material.

5.2 Settings

Metrics. For node evaluation, we follow a retrieval-based protocol [20, 59]. A predicted node is considered correct if it has a non-zero 3D IoU with a ground-truth node and the corresponding ground-truth label ranks within the top-K candidates based on cosine similarity between CLIP [33] embeddings of predicted and ground-truth labels. Node recall is computed as the proportion of ground-truth nodes retrieved under this criterion.

For edge evaluation, a correct triplet (`<node-edge-node>`) is counted as retrieved within top-K only if all its components are individually retrieved within top-K under the same node retrieval criterion, and the relation is matched via cosine similarity of phrase embeddings. We utilize Sentence-BERT [36] because the previous method [20, 59] based on BERT [9] lacked the capability to properly evaluate relational phrases. Affordance evaluation is also performed using Sentence-BERT embeddings, where recall is computed solely based on affordance phrase similarity, independent of node label. We additionally report recall for the associated node, which considers only node pairs without a relation label.

5.3 Quantitative Results

We compare our OP3DSG with representative open-vocabulary 3DSG generation models [14,59]. As neither baseline is originally designed to construct unified 3DSGs, we extend ConceptGraph [14] with additional part-level and functional relation inference via LLM prompts, and augment OpenFunGraph [59] with spatial relation reasoning using the same prompting strategy. Affordance inference is incorporated into both baselines. For fair and up-to-date comparisons, we report results using both the originally adopted GPT-4 [1] (GPT-4-0613) and the more recent GPT-5 [40] (GPT-5-2025-08-07) model, which serves as the LLM backbone of OP3DSG. We also include FunGraph [39] and KeySG [53] in the comparison, both of which rely only on inter-/intra-object relations without explicitly modeling functional relations.

Table 2: Node evaluation on UniGraph3D dataset. The sign $*$ denotes that an additional VLM/LLM prompt is used to infer the part node; ++ indicates the use of the recent GPT-5 [40] model for fair comparison.

Model	Object Node		Part Node		Overall Node		Affordance	
	R@3	R@5	R@3	R@5	R@3	R@5	R@3	R@5
ConceptGraph* [14]	45.4	51.8	17.6	21.5	28.1	32.9	2.3	3.8
ConceptGraph*++ [14]	53.4	61.4	16.7	19.6	30.5	35.3	4.9	5.2
FunGraph [39]	68.5	73.7	28.3	39.1	43.5	52.2	40.3	40.8
FunGraph++ [39]	68.9	73.7	28.5	40.3	43.8	52.9	40.4	40.8
OpenFunGraph [59]	56.2	64.5	49.3	51.0	51.9	56.1	13.7	16.1
OpenFunGraph++ [59]	62.2	69.3	51.7	52.4	55.6	58.8	32.1	33.9
KeySG [53]	74.1	76.5	52.4	54.1	60.6	62.6	25.8	30.5
OP3DSG (Ours)	**84.9**	**93.2**	**83.6**	**86.2**	**84.1**	**88.9**	**76.7**	**83.2**

Table 3: Edge evaluation on UniGraph3D dataset. The sign ‡ denotes that an additional VLM/LLM prompt is used to infer the spatial relation. Assoc. Node refers to the node association metric.

Model	Assoc. Node			Func. Triplet			Spat. Triplet		
	R@3	R@5	R@10	R@3	R@5	R@10	R@3	R@5	R@10
ConceptGraph* [14]	4.2	5.2	7.3	5.9	6.6	8.5	1.7	1.7	2.2
ConceptGraph*++ [14]	7.7	8.7	14.5	11.5	12.7	19.8	2.8	3.1	7.3
FunGraph [39]	9.6	13.3	21.7	0.0	0.0	0.0	21.0	29.1	47.6
FunGraph++ [39]	9.8	13.9	21.9	0.0	0.0	0.0	21.6	30.5	47.9
OpenFunGraph‡ [59]	38.2	41.9	44.5	36.9	40.2	45.2	35.9	39.8	43.1
OpenFunGraph‡++ [59]	45.1	48.7	50.9	41.9	45.9	51.8	44.5	48.2	49.9
KeySG [53]	12.9	17.1	21.5	0.0	0.0	0.0	28.3	37.5	47.1
OP3DSG (Ours)	**53.1**	**66.9**	**76.1**	**51.1**	**64.5**	**74.8**	**52.4**	**66.9**	**77.3**

Node Evaluation. Table 2 presents the results of the node recall on the unified 3DSG generation task. Our framework substantially enhances the overall node performance, with particular gains in part-level recognition, surpassing the strongest baseline by +31.2 points (R@3) and +32.1 points (R@5). This large margin is not merely a byproduct of stronger language models, but reveals a structural limitation of existing open-vocabulary baselines. While upgrading the VLM/LLM backbone moderately improves object-node recall, part-node performance either remains low—indicating that scaling foundation models alone cannot resolve the intrinsic difficulty of open-vocabulary part grounding. In particular, part-centric prompts often collapse into object-level localization, yielding confident object detections but weak or fragmented part representations. In contrast, OP3DSG explicitly tackles this bottleneck at the perception and fusion levels. As a result, parts are not treated as noisy in object detection, but as geometrically stable and semantically coherent entities in the reconstructed scene.

We further evaluate the affordance predictions for each node in Table 2 and observe that OP3DSG significantly outperforms all baselines, achieving 76.7% (R@3) and 83.2% (R@5). Since affordance inference is conditioned primarily on the predicted object/part labels and relies on the LLM's pretrained knowledge, the improved part-node recognition and stable object–part attachment

Table 4: Node Localization Sensitivity study on UniGraph3D dataset.

Model	Object Node		Part Node		Overall Node	
	$IoU_{\geq 0.10}$	$IoP_{\geq 0.25}$	$IoU_{\geq 0.10}$	$IoP_{\geq 0.25}$	$IoU_{\geq 0.10}$	$IoP_{\geq 0.25}$
ConceptGraph* [14]	30.7	27.5	4.1	1.2	14.1	11.1
ConceptGraph++* [14]	32.3	28.7	5.6	2.4	15.6	12.3
FunGraph [39]	49.8	57.8	15.7	8.9	28.6	27.4
FunGraph++ [39]	49.0	57.8	14.5	9.7	27.5	27.8
OpenFunGraph [59]	38.2	30.3	15.0	5.8	23.8	15.0
OpenFunGraph++ [59]	43.4	35.9	15.0	6.0	25.7	17.3
KeySG [53]	50.6	59.4	**30.4**	**23.4**	38.0	**37.0**
OP3DSG (Ours)	**63.7**	**61.8**	25.8	13.0	**40.2**	31.4

Table 5: Ablation study on UniGraph3D dataset.

Model	Object Node		Part Node		Func. Triplet		Spat. Triplet		Affordance	
	R@3	R@5	R@3	R@5	R@3	R@5	R@3	R@5	R@3	R@5
OP3DSG (Ours)	**84.9**	**93.2**	**83.6**	**86.2**	**51.1**	**64.5**	**52.4**	**66.9**	**76.7**	**83.2**
w/o knowledge	76.9	86.5	45.9	48.3	15.5	21.2	12.6	18.8	30.7	33.6
w/o color-dist. feat.	84.9	93.2	79.0	81.6	40.2	52.0	41.7	49.9	64.8	70.8
w/o prior graph	84.9	93.2	83.6	86.2	47.1	58.1	47.6	59.1	74.7	80.9
w/o split LLM	84.9	93.2	83.6	86.2	20.7	27.8	19.3	27.7	40.8	45.0

contribute substantially to the performance gains. These results suggest that language-based affordance reasoning becomes substantially more effective without relying on per-node caption generation once correct interaction anchors—well-grounded parts and their associated objects—are established.

Edge Evaluation. Table 3 presents the edge recall of our method and the baselines on the unified 3DSG generation task. Our framework substantially improves overall edge performance, with gains of 9.2%p (R@3) in functional relations and 7.9%p (R@3) in spatial relations compared to the strongest baseline. These improvements persist even when baselines are strengthened with additional prompting and newer LLMs, indicating that the advantage does not stem from stronger language priors alone, but from more effective structural constraints. In OP3DSG, the geometry-initialized prior graph restricts candidate reasoning to structurally plausible neighborhoods, while the verification-gated multi-agent LLM refinement enforces geometric consistency during relation inference. Consequently, the LLM operates as a constraint-aware refiner rather than an unconstrained edge generator, improving both precision by suppressing spurious edges and recall by reducing missed dependencies in a scalable manner.

Node Localization Sensitivity. To analyze geometric localization accuracy, we evaluate node retrieval under strict overlap constraints, as reported in Table 4. Unlike our main retrieval protocol that uses a non-zero 3D IoU as a minimal spatial consistency filter, we set each threshold to 0.10 for IoU and 0.25 for IoP (Intersection over Prediction) [6,53] when computing R@3. IoP measures

the fraction of the predicted 3D bounding box that lies within the ground-truth box, and primarily evaluates over-expansion or leakage of predictions beyond the true object extent. In contrast, IoU emphasizes overall overlap quality and is more sensitive to under-coverage and center misalignment, particularly when only partially observed surfaces are fused. OP3DSG demonstrates strong overall performance under both criteria, indicating that its advantage extends beyond semantic retrieval to stricter geometric grounding. Nevertheless, OP3DSG still shows limited geometric localization accuracy for part nodes, reflecting the intrinsic difficulty of localizing small components in surface-based fusion.

5.4 Ablation Study

Table 5 ablates key components of OP3DSG. Removing the knowledge-guided control of the entity space causes a substantial drop in part retrieval, accompanied by severe degradation in functional/spatial triplets and affordances. This indicates that incomplete part grounding propagates directly to higher-order graph reasoning. Disabling the color-distribution feature results in only a modest decrease in part recall, yet triplet and affordance performance declines more noticeably, suggesting that even minor part-association instability can be amplified at the relational level. Replacing the geometry-initialized prior graph with a flat list of node attributes reduces triplet and affordance recall, demonstrating that structured spatial context facilitates more reliable LLM-based inference even with identical node sets. Assigning all roles to a single LLM without decomposition leads to a sharp decline in relation and affordance, due not only to increased hallucinations but also to context overload from the enlarged input scope. Overall, these ablations confirm that unified 3DSG quality depends not only on node retrieval accuracy, but also on controlled part candidate expansion, stable part association, and structured geometric context for relation and affordance reasoning.

5.5 Qualitative Results

We visualize portions of the unified 3DSG in Figure 4. The graph jointly represents objects, their parts, and spatial or functional relations between them. For example, the oven and blender are connected to component nodes such as knobs, handles, and lids, while relational edges capture interactions like *on*, *near*, and *provide power*, illustrating the fine-grained structure encoded in the unified 3DSG.

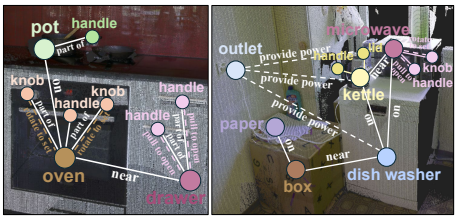

Fig. 4: Qualitative results. We visualize the unified 3DSG in partial regions of two scenes.

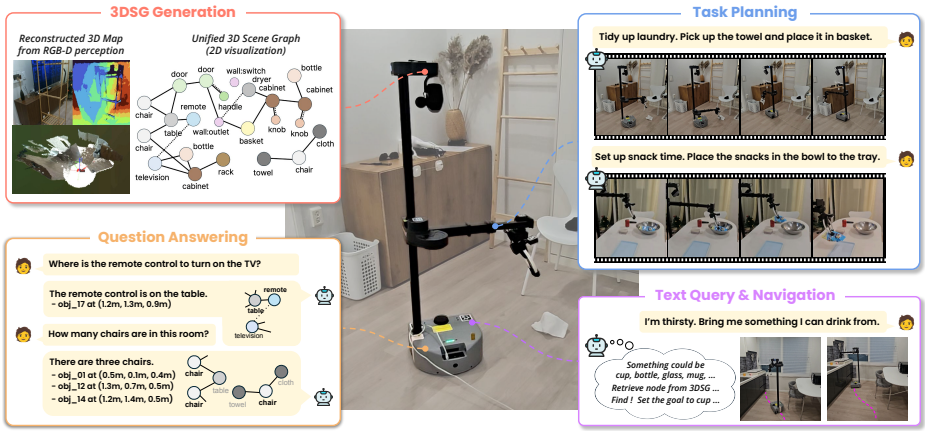

Fig. 5: Robot Applications using OP3DSG. The generated 3DSGs can be used for various downstream tasks, including question answering, task planning, language-conditioned querying, and navigation. See the supplement for implementation details.

6 Real-world Robot Applications

We demonstrate that OP3DSG directly supports diverse real-world embodied tasks (Fig. 5 and demo video). All experiments are conducted on the Stretch3 robot equipped with an Intel RealSense D435i RGB-D camera. During exploration, the robot incrementally reconstructs the environment and generates a unified 3DSG, which serves as a fine-grained representation for various downstream tasks.

(i) ***Question Answering*** demonstrates OP3DSG supports language-guided reasoning over structured scene entities. Given a user query and the unified 3DSG, the LLM performs reasoning directly on the graph and handles questions involving counting, localization, relational inference, and part-level distinction—illustrating the fine-grained structural information in the unified 3DSG.

(ii) ***Task Planning*** demonstrates how OP3DSG grounds high-level language instructions into executable plans. The unified 3DSG maintains persistent object and part nodes with explicit spatial and relational edges, enabling the planner to resolve references to both objects and their components (*e.g.*, handles, buttons). This representation allows instructions to be decomposed into part-aware subgoals and ordered according to spatial dependencies in the graph.

(iii) ***Text Query & Navigation*** demonstrates language-conditioned spatial retrieval with OP3DSG under open-ended user intents. Unlike question answering, queries may not specify explicit object names but instead express functional needs or situational intentions. The system retrieves candidate nodes from the unified 3DSG at both object and part levels and uses the selected node's 3D location to generate a navigation goal.

7 Conclusion

We introduced Unified 3D Scene Graph generation, aiming to represent indoor environments with a single compositional graph that jointly captures objects, parts, spatial/functional relations, and affordances. To make this challenging open-vocabulary setting tractable, we proposed OP3DSG, which couples knowledge-guided control of the entity space, part-aware multi-view 3D fusion, and geometry-anchored, verification-gated LLM refinement to constrain relation and affordance inference. We also presented UniGraph3D, enabling systematic evaluation of part-aware perception and multi-level relational reasoning. The proposed framework achieves substantial gains over prior open-vocabulary baselines. We further demonstrate that OP3DSG can serve as a robust perception backbone for real-world robotic applications, including complex text querying, navigation, and task planning. We expect our formulation, benchmark, and method will facilitate future research on scalable, part-aware 3D scene understanding and its deployment in real-world embodied systems.

Acknowledgements

This research was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220907); by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1C1C1009989); and by the National Research Council of Science & Technology (NST) grant funded by the Korea government (MSIT) (No. GTL25041-000).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023)
2. Agia, C., Jatavallabhula, K.M., Khodeir, M.N.M., Mikvs'ik, O., Vineet, V., Mukadam, M., Paull, L., Shkurti, F.: Taskography: Evaluating robot task planning over large 3d scene graphs. In: Conference on Robot Learning (2022)
3. 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: Int. Conf. Comput. Vis. pp. 5664–5673 (2019)
4. Chang, H., Boyalakuntla, K., Lu, S., Cai, S., Jing, E., Keskar, S., Geng, S., Abbas, A., Zhou, L., Bekris, K., et al.: Context-aware entity grounding with open-vocabulary 3d scene graphs. In: Conference on Robot Learning. pp. 1950–1974 (2023)
5. Chang, Y., Hughes, N., Ray, A., Carlone, L.: Hydra-multi: Collaborative online construction of 3d scene graphs with multi-robot teams. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10995–11002. IEEE (2023)

6. Chen, H., Yang, J., Vondrick, C., Mao, C.: Invite: Interpret and control vision-language models with text explanations. In: Int. Conf. Learn. Represent. (2024)
7. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Coloring channel representations for visual tracking. In: Scandinavian conference on image analysis. pp. 117–129. Springer (2015)
8. Delitzas, A., Takmaz, A., Tombari, F., Sumner, R., Pollefeys, M., Engelmann, F.: Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14531–14542 (2024)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv **abs/1810.04805** (2018)
10. Everts, I., Van Gemert, J.C., Gevers, T.: Evaluation of color stips for human action recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2850–2857 (2013)
11. Feng, M., Hou, H., Zhang, L., Wu, Z., Guo, Y., Mian, A.: 3d spatial multimodal knowledge accumulation for scene graph prediction in point cloud. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9182–9191 (2023)
12. Gadre, S.Y., Ehsani, K., Song, S., Mottaghi, R.: Continuous scene representations for embodied ai. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14849–14859 (2022)
13. Geng, H., Xu, H., Zhao, C., Xu, C., Yi, L., Huang, S., Wang, H.: Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7081–7091 (2023)
14. 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: IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028 (2024)
15. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5356–5364 (2019)
16. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: Partimagenet: A large, high-quality dataset of parts. In: Eur. Conf. Comput. Vis. pp. 128–145. Springer (2022)
17. Hou, H.Y., Lee, C.Y., Sonogashira, M., Kawanishi, Y.: Fross: Faster-than-real-time online 3d semantic scene graph generation from rgb-d images. In: Int. Conf. Comput. Vis. pp. 28818–28827 (2025)
18. Hughes, N., Chang, Y., Carlone, L.: Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. In: Robotics: Science and Systems XVIII (2022)
19. 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)
20. 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: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14183–14193 (2024)
21. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10965–10975 (2022)
22. Li, X., Xu, S., Yang, Y., Cheng, G., Tong, Y., Tao, D.: Panoptic-partformer: Learning a unified model for panoptic part segmentation. In: Eur. Conf. Comput. Vis. pp. 729–747. Springer (2022)

23. Li, Y., Ma, Y., Huo, X., Wu, X.: Remote object navigation for service robots using hierarchical knowledge graph in human-centered environments. Intelligent Service Robotics **15**(4), 459–473 (2022)
24. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. IEEE Trans. Image Process. **24**(12), 5630–5644 (2015)
25. Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J.: Learning object-language alignments for open-vocabulary object detection. In: Int. Conf. Learn. Represent. (2023)
26. Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F.M., Su, H.: Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21736–21746 (2023)
27. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: Eur. Conf. Comput. Vis. pp. 38–55. Springer (2024)
28. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. In: Eur. Conf. Comput. Vis. (2020)
29. Ma, Z., Yue, Y., Gkioxari, G.: Find any part in 3d. In: Int. Conf. Comput. Vis. pp. 7818–7827 (2025)
30. Meletis, P., Wen, X., Lu, C., De Geus, D., Dubbelman, G.: Cityscapes-panoptic-parts and pascal-panoptic-parts datasets for scene understanding. ArXiv **abs/2004.07944** (2020)
31. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: Eur. Conf. Comput. Vis. pp. 728–755. Springer (2022)
32. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 909–918 (2019)
33. 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: Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
34. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7141–7151 (2023)
35. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I.D., Sünderhauf, N.: Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In: Conference on Robot Learning (2023)
36. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. ArXiv **abs/1908.10084** (2019)
37. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. ArXiv **abs/2401.14159** (2024)
38. 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**, 1510 – 1546 (2021)
39. Rotondi, D., Scaparro, F., Blum, H., Arras, K.O.: Fungraph: Functionality aware 3d scene graphs for language-prompted scene interaction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4083–4090. IEEE (2025)

40. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. ArXiv **abs/2601.03267** (2025)
41. Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., Yan, Z.: Going denser with open-vocabulary part segmentation. In: Int. Conf. Comput. Vis. pp. 15453–15465 (2023)
42. Swain, M.J., Ballard, D.H.: Color indexing. International journal of computer vision **7**(1), 11–32 (1991)
43. Szymańska, E., Dusmanu, M., Buurlage, J.W., Rad, M., Pollefeys, M.: Space3d-bench: Spatial 3d question answering benchmark. In: Eur. Conf. Comput. Vis. pp. 68–85. Springer (2024)
44. Van De Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1582–1596 (2009)
45. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Trans. Image Process. **18**(7), 1512–1523 (2009)
46. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3961–3970 (2020)
47. Wang, J., Chakraborty, R., Yu, S.X.: Transformer for 3d point clouds. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 4419–4431 (2019)
48. Wang, L., Li, X., Fang, Y.: Few-shot learning of part-specific probability space for 3d shape segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4504–4513 (2020)
49. Wang, P., Shen, X., Lin, Z.L., Cohen, S.D., Price, B.L., Yuille, A.L.: Joint object and part segmentation using deep learned potentials. In: Int. Conf. Comput. Vis. pp. 1573–1581 (2015)
50. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 24824–24837 (2022)
51. Wei, M., Yue, X., Zhang, W., Kong, S., Liu, X., Pang, J.: Ov-parts: Towards open-vocabulary part segmentation. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 70094–70114 (2023)
52. 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)
53. Werby, A., Rotondi, D., Scaparro, F., Arras, K.O.: Keysg: Hierarchical keyframe-based 3d scene graphs. In: IEEE International Conference on Robotics and Automation (ICRA) (2026)
54. Wu, S.C., Tateno, K., Navab, N., Tombari, F.: Incremental 3d semantic scene graph prediction from rgb sequences. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5064–5074 (2023)
55. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7515–7525 (2021)
56. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15254–15264 (2023)
57. Yang, Y.N., Huang, Y., Guo, Y.C., Lu, L., Wu, X., Lam, E.Y., Cao, Y.P., Liu, X.: Sampart3d: Segment any part in 3d objects. ArXiv **abs/2411.07184** (2024)

58. Yu, F., Liu, K., Zhang, Y., Zhu, C., Xu, K.: Partnet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9483–9492 (2019)
59. 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: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19401–19413 (2025)
60. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1724–1732 (2024)
61. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16793–16803 (2022)
62. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International journal of computer vision **127**(3), 302–321 (2019)
63. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: Eur. Conf. Comput. Vis. pp. 350–368. Springer (2022)
64. Zhou, Y., Gu, J., Li, X., Liu, M., Fang, Y., Su, H.: Partslip++: Enhancing low-shot 3d part segmentation via multi-view instance segmentation and maximum likelihood estimation. ArXiv **abs/2312.03015** (2023)