

Hierarchical and Holistic Open-Vocabulary Functional 3D Scene Graphs for Indoor Spaces

Xinggang Hu^{1,2*}, Chenyangguang Zhang^{3*}, Alexandros Delitzas^{3,4},
Xiangkui Zhang², Marc Pollefeys^{3,5}, Francis Engelmann⁶, and Xiangyang Ji^{1†}

¹Tsinghua University, China ²Dalian University of Technology, China
³ETH Zurich, Switzerland ⁴MPI for Informatics, Germany
⁵Microsoft, Switzerland ⁶USI Lugano, Switzerland

Abstract. Functional 3D scene graphs offer a versatile and flexible representation for 3D scene understanding and robotic manipulation, defined by object nodes, interactive elements, and functional relationship edges. However, their potential remains underexplored due to the limited coverage of existing benchmarks and the overly straightforward design of previous pipelines, which primarily focus on large-scale furniture but lack of hierarchical structures. Therefore, in this work, we extend the benchmark coverage by introducing dense tabletop objects and explicit multi-level functional relationships. This expansion introduces critical challenges involving small-scale, dense, and similar instances, with lack of visual anchoring in relational reasoning, instance confusion during cross-frame fusion, and attribution uncertainty under dynamic viewpoints. To address these issues, we propose an open-vocabulary pipeline based on 2D visual grounding and 3D graph optimization. Specifically, we anchor fine-grained functional edges from 2D visual evidence, and associate nodes across frames in 3D using multiple cues. Furthermore, edge association is formulated as temporal graph optimization, integrating evidence accumulation, entropy regularization, and temporal smoothing to robustly determine the functional connections of each node. Finally, global hierarchy shaping is performed to recover the hierarchical graph structure. Extensive experiments demonstrate that the proposed method can reliably infer functional 3D scene graphs in challenging real-world scenes, thereby further unlocking their potential for practical applications. Code is available at <https://github.com/Hbelief1998/HHOpenFunGraph-ECCV26>.

Keywords: 3D Scene Graph · Scene Understanding · Open Vocabulary

1 Introduction

As virtual reality, robotics, and embodied intelligence continue to advance, it becomes increasingly vital to develop a versatile and accurate 3D scene representation that bridges the gap between high-level 3D understanding and downstream tasks such as manipulation and content generation. Traditionally, 3D

*Equal contribution. [†]Corresponding author.

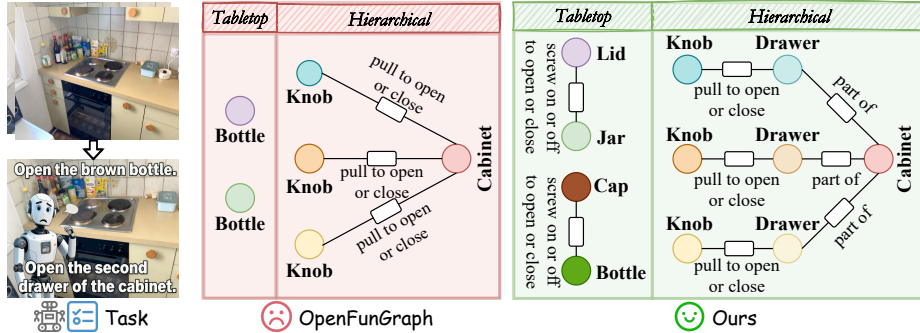


Fig. 1: Hierarchical and holistic functional 3D scene graphs. In contrast to prior approaches [75], we model tabletop manipulable objects and explicit hierarchical object-part structures in functional 3D scene graphs.

scene graphs [2, 6, 19, 32, 33, 50, 51, 60, 66] have been proposed as discrete graph-based representations of 3D environments, consisting of object nodes and their spatial relationships. While such representations have proven effective for robotic navigation and scene understanding, their coarse granularity limits their applicability to fine-grained tasks such as manipulation.

Recently, OpenFunGraph [75] introduced the concept of functional 3D scene graphs by extending traditional scene graphs to include objects, interactive elements, and functional relationships. Although this representation is theoretically flexible and capable of encoding rich semantic and relational properties, its benchmark and pipeline remain limited. The existing setup cannot handle the complexity commonly observed in real-world scenarios. Specifically, the annotations in SceneFun3D [11] and FunGraph3D [75] consider only large furniture items (*e.g.*, cabinets, bathtubs, wardrobes) and lack the tabletop objects that are more common in robotic manipulation tasks. In addition, the intrinsic partonomy and hierarchical structure of objects and parts are not modeled, restricting the applicability of the representation to fine-grained reasoning. For example, as Figure 1 shows, multiple knobs are ambiguously linked to the same cabinet without distinction between drawers, leading to confusion in manipulation queries (*e.g.*, open the second drawer of the cabinet).

To overcome these limitations, we extend the capabilities of the functional 3D scene graph benchmark by introducing: (1) diverse tabletop objects and their associated functional attributes (*e.g.*, bottle-cap, kettle-handle), covering small, manipulable targets commonly found in real-world operations; (2) explicit hierarchical relationships between objects and their constituent parts, treating intermediate structures as functional carriers (*e.g.*, control panel, drawer, lid), and end-effector components that can be directly manipulated as interactive units (*e.g.*, switch, handle, knob), thereby characterizing typical multi-level structures such as oven-panel-switch, cabinet-drawer-handle, and pot-lid-knob.

In real-world scenarios, introducing tabletop objects and hierarchical structures is crucial for downstream applications. However, this benchmark expansion

introduces perception and reasoning challenges that the existing paradigm of OpenFunGraph [75] struggles to address: (1) Lack of Visual Anchoring in Relational Reasoning. The 3D geometry of tiny tabletop parts (*e.g.*, bottle caps and kettle handles) is highly susceptible to reconstruction noise, while hierarchical structures introduce a massive number of fine-grained nodes of the same category in close spatial proximity (*e.g.*, drawers), resulting in severe 3D bounding box aliasing. Despite such geometric degradation, OpenFunGraph still feeds these distorted 3D bounding boxes in natural language format for an LLM to infer relationships [75]. The compounding of unreliable 3D information with the inherent deficiency of the LLM in reasoning over numerical spaces leads to severe error accumulation, rendering it difficult to robustly establish accurate functional connections in the absence of 2D visual evidence. (2) Instance Confusion in Cross-Frame Fusion. In tabletop and hierarchical scenes, dense instances of the same category are prone to boundary and feature overlap during cross-frame fusion (*e.g.*, compactly arranged spice bottles in a kitchen). OpenFunGraph primarily relies on geometric proximity and performs average aggregation of CLIP similarities [75], which easily leads to instance aliasing. (3) Attribution Uncertainty. Tabletop scenes feature dense arrangements of similar instances with mutual occlusions, while fine-grained nodes introduced by hierarchical structures often lack sufficient distinctiveness from certain viewpoints (*e.g.*, blurred boundaries between cabinet drawers and doors). These issues cause the attribution of interaction units to be highly uncertain among similar neighbors, frequently flipping with viewpoint changes. Relying solely on instantaneous scoring like OpenFunGraph makes edge association prone to jitter and erroneous connections.

To overcome the challenges introduced by the tabletop and hierarchical settings in the expanded benchmark, we propose an open-vocabulary pipeline based on direct visual grounding and graph optimization for constructing hierarchical and holistic functional 3D scene graphs, as illustrated in Figure 1. First, to address the lack of visual anchoring in relational reasoning, guided by an inter-actability map constructed via LLMs and open-vocabulary detection, we utilize visual foundation models to generate candidates for multi-level nodes. By fusing 2D geometry with visual-semantic consistency, we infer 2D functional edges from object nodes to fine-grained nodes based on direct 2D visual evidence. Second, to mitigate instance confusion during cross-frame fusion, we introduce a multi-cue node association mechanism. This integrates appearance, semantics, 2D projection, and 3D geometric information to suppress erroneous associations among dense instances of the same category, thereby forming stable cross-frame node trajectories. Third, regarding attribution uncertainty under dynamic viewpoints, we model edge association as time-evolving variables within a graph optimization framework. Incorporating historical evidence accumulation, entropy regularization, and temporal smoothing, this optimization robustly determines edge connections amidst candidate competition caused by viewpoint changes and observation noise, effectively preventing association jitter and erroneous connections. Finally, a hierarchy completion step refines the structure by optimizing semantic and geometric consistency, resulting in the final hierarchical graph.

Experiments on our expanded benchmark demonstrate our method’s effectiveness in complex scenes featuring tabletop objects and hierarchical structures, outperforming the FunGraph [52] and modified OpenFunGraph [75] baselines by up to 13.9% and 20.5% in node and edge triplet retrieval, respectively.

2 Related Works

3D Scene Understanding. 3D indoor scene understanding has long focused on closed-set 3D semantic/instance segmentation [3, 9, 14, 20, 21, 24–26, 30, 34, 36, 45, 46, 53, 57–59, 62, 63, 72]. Subsequent work, driven by vision-language foundation models, has extended to open-vocabulary 3D semantic segmentation [15, 29, 31, 44, 47, 55, 56, 70, 81, 82] and 3D vision-language alignment [5, 10, 22, 27, 43, 49, 67, 79]. Recent part-level open-vocabulary methods further recognize fine-grained object parts from open textual descriptions [54, 64], highlighting the need to move beyond whole-object perception. However, they mainly focus on part segmentation or recognition, rather than temporally fused 3D functional entities and graph-structured functional relationships. Unlike these perception-oriented works, SceneFun3D [11] proposed a benchmark containing interactive elements. Building on this, FunGraph3D [75] further provides object annotations and the relationships between interactive elements and objects. However, this benchmark does not yet consider tabletop manipulable objects or the hierarchical structure between objects and parts. This paper extends this benchmark by introducing more holistic tabletop objects and hierarchical annotations.

Affordance Understanding. Affordance understanding concerns how an environment can be operated upon. Existing methods predominantly predict affordance heatmaps from images [13, 73], videos [17, 40, 71], or 3D representations [4, 7, 8, 16, 68, 69, 76, 77], or learn interaction strategies from human-scene interaction demonstrations [4, 7, 8, 12, 16, 68, 69, 76, 77]. However, existing approaches typically attach affordances to the object holistically, merely addressing “how a certain object can be used,” failing to specify the concrete operational parts. In contrast, we explicitly bind affordances to specific interactive parts, uniformly modeling hierarchical operational relationships via functional edges.

3D Scene Graphs. 3D scene graphs organize indoor entities and their spatial/semantic relationships in a node-edge format [2, 32, 50, 51, 55, 60, 61, 65, 66, 74, 78]. Building on this, functional 3D scene graphs further incorporate interactive elements as nodes, modeling the functional connections between objects and interactive elements. IFR-Explore [35] mines functional relationships in synthetic scenes using reinforcement learning, but is limited by a closed-set setting and ignores part-level elements; open-vocabulary 3D scene graph methods such as Open3DSG [33] and ConceptGraph [19] primarily remain at the object-level, addressing only object nodes and a few spatial relationships. Furthermore, FunGraph [52] introduces functional part nodes and **has-part** edges into the 3D scene graph through 2D functional part detection and VLM annotation, OpenFunGraph [75] builds upon this by employing open-vocabulary object and interactive element detection and sequentially inferring functional relationships.

However, both have limited coverage of tabletop manipulable objects and do not explicitly model the hierarchical object-part structure. To address the aforementioned limitations, this paper constructs a functional 3D scene graph that simultaneously possesses a multi-level structure and broader coverage of manipulable objects, through 2D functional relationship anchoring, cross-frame node association, functional graph optimization, and global graph refinement.

3 Problem Definition

This work aims to address the task of automatically constructing a hierarchical functional 3D scene graph $\mathcal{G}$ for holistic items in general indoor scenes given a sequence of posed RGB-D images.

The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined with: (1) The node set $\mathcal{V} = \mathcal{O} \cup \mathcal{C} \cup \mathcal{U}$, where Objects $\mathcal{O}$ are the ultimate hosts of functionality and serve as interaction targets (*e.g.*, cabinet, pot, oven, etc.); functional carriers $\mathcal{C}$ are structures on objects that undertake intermediate functions or operable mid-level units (*e.g.*, drawer, lid, control panel, etc.); interactive units $\mathcal{U}$ are the end-effector elements that can be directly manipulated by a user or robot to trigger functions (*e.g.*, handle, knob, switch, etc.). (2) The directed edge set $\mathcal{E}$ represents functional relationships, characterizing the compositional, subordinate, or control interactions among nodes. These relationships constitute a functional hierarchy: depending on the observability and modeling of the functional carrier layer $\mathcal{C}$, two functional relational structures emerge, namely $\mathcal{O} \leftarrow \mathcal{C} \leftarrow \mathcal{U}$ (*e.g.*, oven $\leftarrow$ control panel $\leftarrow$ knob) and $\mathcal{O} \leftarrow \mathcal{U}$ (*e.g.*, bottle $\leftarrow$ cap). The objective of this work is to jointly infer the 3D properties (*e.g.*, position, dimension, and semantics) of objects $\mathcal{O}$, functional carriers $\mathcal{C}$, and interaction trigger units $\mathcal{U}$, alongside their functional relationships $\mathcal{E}$, to ultimately construct a complete scene graph $\mathcal{G}$.

4 Method

To jointly infer the 3D attributes (position, size, semantics, etc.) of objects $\mathcal{O}$, functional carriers $\mathcal{C}$, and interactive units $\mathcal{U}$, as well as the relationships $\mathcal{E}$ between them, thereby constructing the holistic hierarchical functional 3D scene graph $\mathcal{G}$, this study proposes a three-stage progressive reasoning framework, shown in Figure 2. First, in the 2D relation anchoring stage, we construct a dynamic interactability map to guide the open-vocabulary instance detection and segmentation of nodes $\mathcal{V} = \mathcal{O} \cup \mathcal{C} \cup \mathcal{U}$, and generate 2D functional edges based on visual evidence by integrating geometric and visual-semantic consistency evaluations. Next, we associate nodes across frames by fusing appearance, semantics, 2D projection consistency, and 3D geometric information to obtain temporally consistent global node trajectories. The corresponding candidate edges are modeled as time-evolving graph variables, and the functional edges $\mathcal{E}$ are robustly determined using a graph optimization strategy that combines cross-frame evidence accumulation with entropy regularization and temporal smoothing. Finally, in the global graph construction stage, we perform hierarchy shaping on

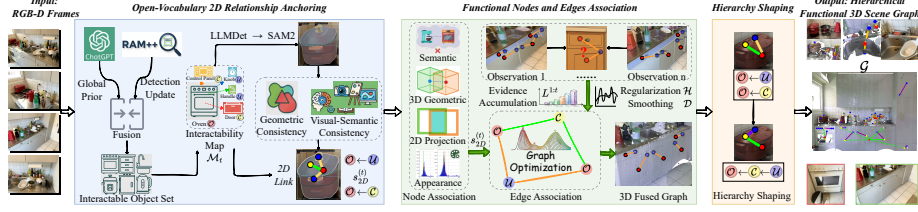


Fig. 2: Overview of the hierarchical functional 3D scene graph construction pipeline.

the associated functional edges to explicitly model the functional hierarchy of $\mathcal{O} \leftarrow \mathcal{C} \leftarrow \mathcal{U}$, ultimately outputting the complete $\mathcal{G}$.

4.1 Open-Vocabulary 2D Relationship Anchoring

The compounding of 3D geometric degradation and the inherent spatial reasoning deficiencies of LLMs leads to severe error accumulation. Specifically, tabletop tiny parts suffer from unstable localization under reconstruction noise, and dense hierarchical nodes (*e.g.*, adjacent drawers) exhibit severe bounding box aliasing. Consequently, 3D spatial proximity becomes non-discriminative. To circumvent this issue, we abandon the paradigm of relying on distorted 3D coordinates for inference and instead directly extract visual evidence from 2D frames. This stage aims to generate reliable functional edge candidates and their associated visual evidence scores, providing an unambiguous basis for subsequent 3D graph fusion.

Generation of Dynamic Interactability Map. To guide open-vocabulary instance detection and relationship reasoning, we introduce a Dynamic Interactability Map $\mathcal{M}_t$, which specifies the manipulable objects $\mathcal{O}$ identified in the current frame t and their associated functional component ($\mathcal{C}, \mathcal{U}$) types. The construction of $\mathcal{M}_t$ fuses the general knowledge of an LLM with the visual evidence from the current frame. First, we utilize the LLM GPT-5 [42] based on task definitions to generate a prior list of globally manipulable objects (*e.g.*, cabinet, oven, washing machine, kettle, etc.). However, this global prior can hardly exhaust scene-specific fine-grained categories and does not reflect frame-level visibility. Therefore, we subsequently identify instances at frame t using RAM++ [28, 80], filter for manipulability via an LLM, and integrate global priors to determine the set of manipulable objects $\mathcal{O}$ for the current frame. Finally, we prompt the LLM, for the objects in $\mathcal{O}$, to generate their associated functional carrier $\mathcal{C}$ and interactive unit $\mathcal{U}$ types, thus updating $\mathcal{M}_t$ dynamically.

Open-Vocabulary Node Candidate Generation. Guided by $\mathcal{M}_t$, we employ a two-stage pipeline to generate the set of detected nodes for the current frame, denoted as $\mathcal{N}_t = \{n_k\}$. LLMDet [18] generates bounding boxes $\mathbf{b}_{n_k}$, categories c_{n_k} , and confidence scores s_{n_k} conditioning on $\mathcal{M}_t$. Subsequently, we prompt SAM2 [48] with $\mathbf{b}_{n_k}$ to yield masks M_{n_k} and extract appearance features $\mathbf{h}_{n_k}$. Each node n_k comprises attributes including $(\mathbf{b}_{n_k}, c_{n_k}, s_{n_k}, M_{n_k}, \mathbf{h}_{n_k})$.

2D Functional Edge Generation and Scoring. Based on the detected node $\mathcal{N}_t$ and the Dynamic Interactability Map $\mathcal{M}_t$, this stage aims to construct and

score functional edge ($o \leftarrow f$) candidates, where $o \in \mathcal{O}$ is an object node and $f \in \mathcal{C} \cup \mathcal{U}$ is a fine-grained node. Note that we do not directly construct the $\mathcal{O} \leftarrow \mathcal{C} \leftarrow \mathcal{U}$ hierarchy in the 2D stage due to 1) increased combinatorial complexity in candidate generation and 2) high ambiguity under single-frame observations. Instead, it is deferred to the global hierarchy shaping stage (Section 4.3), which utilizes global 3D information after the sequence processing is complete.

The generation of candidate pairs (o, f) should first follow the semantic constraints provided by $\mathcal{M}_t$. Subsequently, we employ a cascaded verification process: candidate pairs must first pass a pre-filtering based on node confidence ($\mathcal{S}_{\text{det}} = \min(s_o, s_f) > \tau_{\text{det}}$) and geometric consistency ($\mathcal{G}_{\text{camc}} > \tau_{\text{geo}}$). The thresholds are set to $\tau_{\text{det}} = 0.25$ and $\tau_{\text{geo}} = 0.90$ to cover as many detection candidates as possible, while completing the filtering through strict geometric constraints, thus balancing comprehensiveness and accuracy. Here, $\mathcal{G}_{\text{camc}} = \frac{|M_f \cap (\hat{M}_o \oplus \Delta)|}{|M_f|}$ ensures that the mask M_f is contained within the mask $\hat{M}_o$ ($\oplus \Delta$ denotes a pixel-level morphological dilation on the object mask to tolerate contour errors). Candidate pairs ($o \leftarrow f$) that pass the pre-filtering are then subjected to a final visual-semantic consistency assessment by the VLM LLAVA v1.6 [37–39] to generate the 2D functional edge score $s_{2D}^{(t)}(o \leftarrow f) \in (0, 1]$. $s_{2D}^{(t)}$ quantifies the support of the current frame for this ($o \leftarrow f$) connection and will serve as a key input for the 3D spatiotemporal evidence accumulation.

4.2 Functional Nodes and Edges Association

Although reliable single-frame 2D evidence is extracted in the previous stage, lifting it to 3D introduces critical challenges arising from tabletop and hierarchical settings: cross-frame instance confusion and attribution uncertainty. The small-scale, densely arranged nodes with prone-to-overlap boundaries (*e.g.*, bottles, drawers) inherent to tabletop and hierarchical structures often lead to erroneous cross-frame associations. Furthermore, single-frame functional edges are unstable due to occlusions and viewpoint changes, causing fine-grained nodes to exhibit attribution competition among multiple neighboring candidates (*e.g.*, handle attribution at the junction of cabinet opening structures), making edge states highly susceptible to noise-induced jitter. To address this, we first robustly associate nodes via multi-cue fusion, and subsequently model functional edges as time-evolving graph variables. By jointly constraining long-term evidence accumulation, entropy regularization, and temporal smoothing, we robustly determine functional connections within a graph optimization framework.

Node Association. In frame t , to associate the set of detected nodes with 2D functional edges $\mathcal{N}_t = \{n_k\}$ with the set of node instances in the map $\mathcal{V} = \{v_j\}$, we define a bipartite graph edge weight $S^{(t)}(n_k \rightarrow v_j)$:

$$\begin{aligned} S^{(t)} = & w_{\text{iou}} \text{IoU}(\mathbf{b}_{n_k}, \Pi_t(\mathcal{P}_{v_j})) + w_{\text{geo}} \mathcal{K}(\|\mathbf{x}_{n_k}^{3D} - \boldsymbol{\mu}_{v_j}\|) \\ & + w_{\text{app}} \cos(\mathbf{h}_{n_k}, \mathbf{H}_{v_j}) + w_{\text{sem}} \mathcal{J}(c_{n_k}, c_{v_j}). \end{aligned} \quad (1)$$

Here, $\mathcal{P}_{v_j}$ denotes the accumulated 3D point cloud of the map node v_j , $\Pi_t(\mathcal{P}_{v_j})$ denotes its projected 2D bounding box in frame t , and $\mathbf{H}_{v_j}$ denotes its ac-

cumulated appearance histogram updated over time. We set $w_{\text{iou}} = 0.5$ and $w_{\text{geo}} = w_{\text{app}} = w_{\text{sem}} = 0.5/3$. This configuration prioritizes 2D projection consistency (w_{iou}) to ensure stable tracking under viewpoint changes, while leveraging geometric anchors, appearance discrimination, and semantic compatibility cues in a balanced manner to address complex scenarios such as dense occlusion, class confusion, and label noise. Specifically, IoU measures 2D projection consistency by comparing the detection box $\mathbf{b}_{n_k}$ with the map node’s projected box $\Pi_t(\mathcal{P}_{v_j})$. $\boldsymbol{\mu}_{v_j}$ represents the centroid of the accumulated point cloud for map node v_j , serving as a stable anchor for geometric association. $\mathcal{K}$ denotes a Gaussian kernel $\mathcal{K}(d) = \exp(-d^2/2\sigma^2)$ applied to the Euclidean distance $d = \|\mathbf{x}_{n_k}^{3D} - \boldsymbol{\mu}_{v_j}\|$, mapping the distance to a similarity score within $(0, 1]$. The appearance similarity $\cos(\mathbf{h}_{n_k}, \mathbf{H}_{v_j})$ compares the color histogram of the current frame with the node’s accumulated histogram, enabling effective discrimination even when objects are semantically identical and spatially proximate (*e.g.*, tightly arranged bottles of the same category) [23, 41]. Semantic compatibility $\mathcal{J}$ is defined as the cosine similarity between CLIP features, facilitating the association of synonymous expressions in an open-vocabulary setting (*e.g.*, “handle” vs. “drawer handle”).

To improve matching efficiency and prune false candidates, we only build graph edges between candidate pairs (n_k, v_j) that satisfy the gating conditions: $\|\mathbf{x}_{n_k}^{3D} - \boldsymbol{\mu}_{v_j}\| < \tau_d \wedge \text{IoU}(\mathbf{b}_{n_k}, \Pi_t(\mathcal{P}_{v_j})) > 0 \wedge S^{(t)}(n_k \rightarrow v_j) \geq \tau_{\text{ass}}$. Here, $\tau_d = \min(0.15 \text{ m}, 0.5 \cdot d_{v_j})$, with d_{v_j} being the diagonal length of the bounding box of v_j , to achieve a scale-adaptive geometric proximity constraint; $\tau_{\text{ass}} = 0.45$ filters low-quality matches. On this sparse bipartite graph, we solve the maximum-weight one-to-one matching problem:

$$\max_{\mathbf{A}} \sum_{k,j} A_{kj} S^{(t)}(n_k \rightarrow v_j) \quad \text{s.t.} \quad A_{kj} \in \{0, 1\}, \quad \sum_j A_{kj} \leq 1, \quad \sum_k A_{kj} \leq 1. \quad (2)$$

This matching is non-exhaustive: due to the gating, some nodes may remain unmatched. Successfully matched pairs (n_k, v_j) are used to update the state of the map node instance v_j (*e.g.*, fusing the new observation point cloud into $\mathcal{P}$, recomputing the centroid $\boldsymbol{\mu}$, and updating the appearance $\mathbf{H}$ via exponential moving average); unmatched detections n_k are initialized as new map nodes.

Edge Association. The frame-by-frame evidence provided by the 2D layer is sparse, noisy, and lacks temporal consistency. Specifically, under the expanded benchmark, manipulable objects in tabletop scenes exhibit minute scales, and nodes of the same category coexist in close proximity alongside occlusions. In hierarchical scenes, fine-grained nodes (*e.g.*, adjacent drawer panels) not only suffer from severe appearance homogenization, but their physical boundaries are also difficult to clearly define across various viewpoints. These issues pose significant challenges for determining functional edges within the map: 1) Single-frame visual and geometric cues are highly sensitive to occlusion and viewpoint, rendering hard assignments based on instantaneous scores susceptible to sporadic noise; 2) Fine-grained nodes exhibit multi-candidate ambiguity when facing similar objects (*e.g.*, side-by-side drawers, bottles), and lacking uncertainty modeling, early noise can mislead and lock in erroneous associations; 3) Perturbations

in instantaneous evidence can cause frequent association switching, destabilizing the functional topology. To address these issues, we model edge association as a time-evolving graph optimization problem, robustly determining functional edges under multi-candidate ambiguity and noisy observations through joint design of evidence accumulation, entropy regularization, and temporal smoothing.

Specifically, for an $f \in \mathcal{C} \cup \mathcal{U}$, we first construct its object candidate set $\mathcal{O}_f$ based on 2D functional edge candidates and valid 3D reconstruction conditions. Let $z_{o \leftarrow f}^{(t)}$ denote the soft assignment probability of associating f with a candidate object $o \in \mathcal{O}_f$ at time t , and let $\mathbf{z}_f^{(t)}$ be the distribution over $\mathcal{O}_f$, satisfying $\sum_o z_{o \leftarrow f}^{(t)} = 1$ and $z_{o \leftarrow f}^{(t)} \geq 0$. We optimize the soft assignment distributions for all fine-grained nodes at each time t by maximizing the objective function:

$$\max_{\mathbf{z}_f^{(t)}} \left[\sum_{o \in \mathcal{O}_f} z_{o \leftarrow f}^{(t)} L^{1:t}(o \leftarrow f) + \mathcal{H}(\mathbf{z}_f^{(t)}) - \mathcal{D}(\mathbf{z}_f^{(t)}, \mathbf{z}_f^{(t-1)}) \right]. \quad (3)$$

This objective function balances three aspects. The first term is the data fidelity term, at the core of which is the historically accumulated evidence $L^{1:t}$. To address the issues of sparse and noisy evidence, we employ log-odds accumulation to fuse observations across time:

$$L^{1:t}(o \leftarrow f) := \sum_{\tau \in \mathcal{T}_f^t(o)} \text{logit}(s_{2D}^{(\tau)}(o \leftarrow f)), \quad (4)$$

where $\text{logit}(x) = \log \frac{x}{1-x}$, and $\mathcal{T}_f^t(o)$ is the set of frame indices up to t where the edge $(o \leftarrow f)$ was validly observed (*i.e.*, $o \in \mathcal{O}_f$). This method leverages long-term observations to offset instantaneous noise, widening the evidence gap between strong and weak candidates over time.

The second term is the entropy regularization term $\mathcal{H}(\mathbf{z}_f^{(t)})$, used to suppress premature polarization:

$$\mathcal{H}(\mathbf{z}_f^{(t)}) = - \sum_{o \in \mathcal{O}_f} z_{o \leftarrow f}^{(t)} \log z_{o \leftarrow f}^{(t)}. \quad (5)$$

This term encourages the distribution to maintain uncertainty when evidence is still insufficient, avoiding locking onto a single association when evidence is ambiguous.

The third term is the temporal smoothing term $\mathcal{D}(\cdot, \cdot)$, used to suppress association jitter:

$$\mathcal{D}(\mathbf{z}_f^{(t)}, \mathbf{z}_f^{(t-1)}) = \frac{1}{2} \sum_{o \in \mathcal{O}_f} (z_{o \leftarrow f}^{(t)} - z_{o \leftarrow f}^{(t-1)})^2. \quad (6)$$

This term penalizes drastic changes in the soft assignment distribution between adjacent frames, ensuring the association updates only when new evidence is sufficiently strong.

This objective defines a graph optimization problem over the candidate object set $\mathcal{O}_f$ for the fine-grained node f . As the final determination of a functional edge is essentially a discrete choice from a finite set of candidate objects, we construct an integrated decision score:

$$\Lambda^{(t)}(o \leftarrow f) := L^{1:t}(o \leftarrow f) + \log z_{o \leftarrow f}^{(t)}, \quad (7)$$

and for each f , we take the following over the $\mathcal{O}_f$:

$$\tilde{o}_f^{(t)} := \arg \max_{o \in \mathcal{O}_f} \Lambda^{(t)}(o \leftarrow f), \quad (8)$$

i.e., we adopt a Top-1 strategy to select the object with the highest score, incorporating $(\tilde{o}_f^{(t)} \leftarrow f)$ into $\mathcal{E}$ as the functional edge at time t .

4.3 Global Hierarchy Shaping

As described in Section 4.1, directly parsing the multi-level structure at the 2D layer leads to candidate combinatorial complexity and single-frame view ambiguity. Whereas in the 3D layer, nodes and functional edges have completed cross-frame fusion via graph optimization, 2D noise is significantly suppressed, and reliable evidence can be provided using precise 3D geometry. Therefore, multi-hierarchical relationships are better determined in 3D.

Given the result from the last step $\mathcal{F}_o = \{f \in \mathcal{F} \mid (o \leftarrow f) \in \mathcal{E}\}$ for each object $o \in \mathcal{O}$, the goal of global hierarchy shaping is to rewrite suitable $o \leftarrow u$ edges among them into hierarchical $o \leftarrow c \leftarrow u$ connections. We leverage the hierarchical role feasibility $r_f^{\mathcal{C}}(o), r_f^{\mathcal{U}}(o) \in \{0, 1\}$ provided by $\mathcal{M}_t$ to define the candidate functional carrier set $\mathcal{C}_o = \{f \in \mathcal{F}_o \mid r_f^{\mathcal{C}}(o) = 1\}$ and the candidate interactive unit set $\mathcal{U}_o = \{f \in \mathcal{F}_o \mid r_f^{\mathcal{U}}(o) = 1\}$ within $\mathcal{F}_o$. A one-time pairing is performed on $\mathcal{C}_o \times \mathcal{U}_o$ by defining binary pairing variables $Y_{c \leftarrow u} \in \{0, 1\}$ where $\sum_{c \in \mathcal{C}_o} Y_{c \leftarrow u} \leq 1 \quad (\forall u \in \mathcal{U}_o)$, ensuring that each interactive unit u is assigned to at most one functional carrier c . The pairing objective is to maximize a score based on semantics and geometry:

$$\max_{\{Y_{c \leftarrow u}\}} \sum_{u \in \mathcal{U}_o} \sum_{c \in \mathcal{C}_o} Y_{c \leftarrow u} (C_{\text{prior}}(c, u) + G_{\text{near}}(c, u)). \quad (9)$$

Here, $C_{\text{prior}}(c, u)$ is the semantic compatibility defined by $\mathcal{M}_t$; $G_{\text{near}}(c, u)$ is the 3D geometric proximity score. Finally, we restructure the graph based on the optimization result $Y_{c \leftarrow u}$: if $Y_{c \leftarrow u} = 1$, the $o \leftarrow u$ edge is removed and $c \leftarrow u$ is added to construct the $o \leftarrow c \leftarrow u$ hierarchy; if $\sum_c Y_{c \leftarrow u} = 0$, the direct $o \leftarrow u$ connection is retained.

5 Datasets and Benchmarks

OpenFunGraph [75] provides a coarse-grained benchmark on functional 3D scene graphs on FunGraph3D and modified SceneFun3D, with most annotations fo-

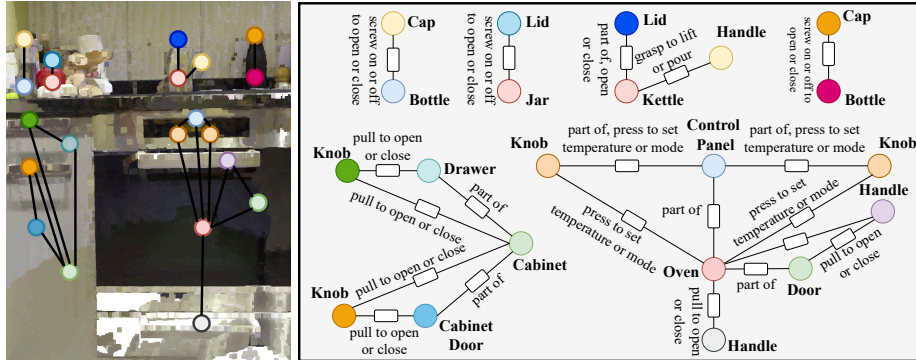


Fig. 3: Examples of the improved benchmark. We introduce tabletop manipulable objects and hierarchical relationships.

cusing on large furniture (*e.g.* cabinet, refrigerators), linking all interactive elements to the object without considering natural hierarchies (*e.g.* wardrobe-drawer-knob). As shown in Figure 3, our improved benchmark makes it more fine-grained and closer to realistic setting by introducing (1) various tabletop objects which play a common role in manipulation tasks (*e.g.* kettle, bottle); (2) hierarchical relations which indicate the specific functional carrier of each interactive element (*e.g.* pot-lid-knob, chest-drawer-knob). To enable it, we mimic the annotation tool in OpenFunGraph and extend it with hierarchical annotation features. Annotators can navigate the 3D scene and annotate the instances of object - functional carrier - interactive unit triplets with a free-form label. To enable the above extensions, we conducted a systematic hierarchical functional annotation and statistical analysis on FunGraph3D [75] and SceneFun3D [11]. FunGraph3D contains 722 nodes ($\mathcal{O} = 224$, $\mathcal{C} = 94$, $\mathcal{U} = 404$), with 592 functional edges in total; among them, 118 are hierarchical relations, and 140 belong to the tabletop category. In contrast, we find that the scenes in SceneFun3D do not include many tabletop items. Therefore, it is just used for evaluating hierarchical performance. It contains 417 nodes ($\mathcal{O} = 106$, $\mathcal{C} = 90$, $\mathcal{U} = 221$), with 438 functional edges; among them, 120 are hierarchical relations.

6 Experiments

6.1 Experimental Setup

Datasets and Metrics. We evaluate functional 3D scene graphs on the extended-annotation SceneFun3D and FunGraph3D datasets. The evaluation protocol follows the retrieval-based setting of OpenFunGraph [75], but with two key modifications to accommodate the semantic granularity differences and dense local label distribution introduced by free-text labels and our tabletop/multi-level annotation extensions. First, in node/triplet evaluation, a **hit** is defined as Recall@3/Recall@5, or a similarity score exceeding 0.75/0.70, to mitigate false neg-

atives caused by subtle phrasing differences in synonymous labels. Second, the Hungarian algorithm is used to perform optimal one-to-one matching between ground-truth nodes and predicted nodes, preventing a single predicted node or edge from being `hit` by multiple close ground-truths, which would inflate the recall rate. Apart from these two points, the rest of the procedure remains consistent with OpenFunGraph. Considering our introduction of tabletop objects and hierarchical structures, node metrics are reported separately for three levels: objects, functional carriers, and interactive units, with the performance on the tabletop subset tallied separately. For triplet evaluation, in addition to overall results, we also report performance on the hierarchical and tabletop subsets.

Baselines. We primarily compare our method against two representative approaches: FunGraph [52] and OpenFunGraph [75]. FunGraph proposed a functional-aware 3D scene graph, establishing `has-part` connections for objects and functional parts in a three-stage pipeline using 2D-trained functional part detection and VLM contextual annotation. OpenFunGraph provides a baseline closer to our task setting for inferring functional 3D scene graphs in an open-vocabulary manner. To avoid performance discrepancies caused by the language model, the GPT-4 [1] used in the original OpenFunGraph was replaced with GPT-5 [42], the model used in this study. However, OpenFunGraph faces difficulties in complex scenarios that involve tabletop objects and hierarchical relationships. Therefore, while maintaining its simple recognize-then-connect paradigm, we incorporate additional prompts into both its detection module and the LLM reasoning process. By retaining tabletop categories in its detection’s semantic filter and providing the LLM with systematic examples of tabletop functions, we implemented *OpenFunGraph Tabletop*. Building on this, we further incorporate relevant examples and instructions for constructing hierarchical relationships into the LLM prompts, resulting in *OpenFunGraph Hierarchy*.

Data processing. Achieving fine-grained scene understanding inevitably leads to a surge in information density. For instance, on the FunGraph3D dataset, the number of nodes detected and recalled by our system is 2.1 and 1.8 times that of OpenFunGraph, respectively. To control computational overhead, we apply keyframe downsampling to all input sequences. This strategy compresses our runtime to be comparable with the baseline. All subsequently reported metrics for our method are based on these downsampled sequences. The ability to maintain an extremely high recall rate under such sparse observations is attributed to the reliable anchoring of 2D functional edges based on concrete visual evidence, alongside the robust multi-cue association and graph optimization mechanisms, whose historical accumulation and temporal smoothing effectively bridge the temporal gaps caused by downsampling.

6.2 Results

Quantitative Comparison. The quantitative results are shown in Table 1 and Table 2. Since SceneFun3D does not contain tabletop objects at the node level, we get a similar node retrieval ability with OpenFunGraph on easily detected large furniture. However, we consistently outperform on functional carriers for

Table 1: Node evaluation on the SceneFun3D and FunGraph3D datasets. “–” indicates SceneFun3D does not contain Tabletop Nodes.

Methods	SceneFun3D [11]					FunGraph3D [75]				
	Objects	Fun. Carriers	Inter. Elements	Overall Nodes		Objects	Fun. Carriers	Inter. Elements	Tabletop Nodes	Overall Nodes
FunGraph	47.2	32.2	55.2	48.2		28.6	26.6	20.3	25.7	23.7
OpenFunGraph	72.6	13.3	68.8	57.8		49.1	11.7	28.7	39.3	32.8
OpenFunGraph Tabletop	-	-	-	-		53.5	11.7	42.1	52.1	41.7
OpenFunGraph Hierarchy	72.6	47.8	68.8	65.2		53.5	41.5	43.3	55.7	46.3
Ours	75.5	66.7	70.1	70.7		56.3	78.7	58.2	71.4	60.2

Table 2: Triplet evaluation on the SceneFun3D and FunGraph3D. “×” indicates the method fails to recall the triplets.

Methods	SceneFun3D [11]		FunGraph3D [75]		
	Hierarchical	Overall	Hierarchical	Tabletop	Overall
FunGraph [52]	×	23.5	×	6.4	19.1
OpenFunGraph [75]	×	30.8	×	13.6	21.8
OpenFunGraph Tabletop	-	-	×	42.1	28.5
OpenFunGraph Hierarchy	12.5	37.6	16.1	42.9	35.1
Ours	52.5	50.9	61.9	61.4	55.6

Table 3: Ablations on FunGraph3D [75] dataset for nodes and triplets.

Methods	Nodes		Triplets		
	Table-top	Overall	Hier.	Table-top	Overall
Assoc-OpenFG	58.1	46.1	46.6	48.6	40.9
No-GO-Count	67.1	55.1	50.8	57.9	47.1
2D-Hierarchy	70.3	51.9	44.9	60.7	49.2
Ours	71.4	60.2	61.9	61.4	55.6

our designated process on that hierarchy. On the FunGraph3D dataset, which features numerous tabletop objects and hierarchical structures, we demonstrate significant advantages across all node evaluation metrics. To be specific in each baseline, FunGraph relies on 2D-trained part detection and contextual annotation, lacking open-vocabulary capabilities, cross-frame fusion, and explicit hierarchical modeling. It struggles to form stable grounding for dense small parts and multi-level relationships, thus exhibiting overall lower node and triplet metrics on both datasets. OpenFunGraph relies solely on geometric proximity and CLIP mean aggregation, making it difficult to distinguish small-scale, textureless tabletop nodes and densely arranged fine-grained hierarchical nodes (*e.g.*, bottles, drawers) during 3D fusion. Overall, the improvement in our node recall is attributed to: 1) the dynamic Interactability Map, which expands the functional node candidate space; 2) more robust cross-frame node association; and 3) building upon 2D functional anchoring, the carrier role is explicitly preserved via temporal association and 3D global hierarchy shaping.

At the triplet level, the significant discrepancy between the Tabletop and Hierarchy subsets corroborates the challenges regarding the lack of visual anchoring and attribution uncertainty. FunGraph [52] relies on detection box overlap and incremental fusion; however, due to the lack of global temporal optimization, it struggles to rectify single-view occlusions, false detections, and accumulated errors. OpenFunGraph falters under noise and aliasing due to the lack of explicit modeling for tabletop and hierarchical structures and its reliance on coarse 3D spatial overlap to filter candidates. OpenFunGraph Hierarchy still adheres to the geometric proximity paradigm, unable to resolve bounding box aliasing among spatially adjacent nodes of the same category (*e.g.*, dense drawers); meanwhile, OpenFunGraph Tabletop is plagued by the geometric drift of tiny nodes and attribution ambiguity of interactive units caused by a lack of temporal consistency. In contrast, we leverage 2D visual anchoring and temporal optimization to robustly determine functional relationships.

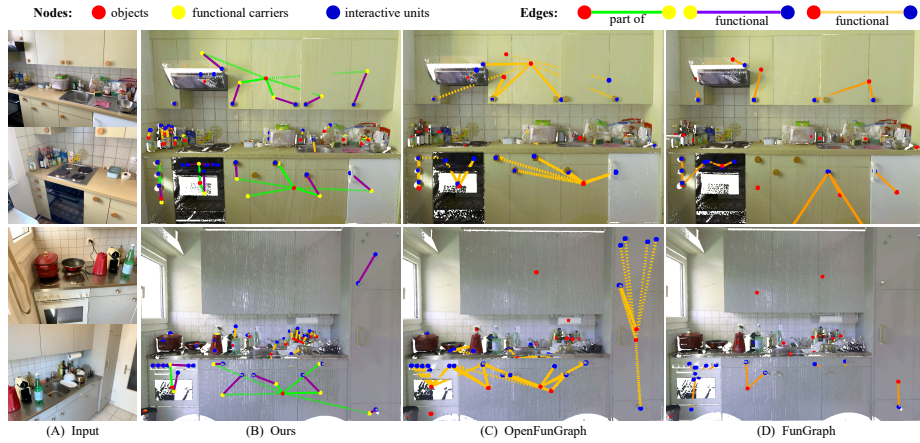


Fig. 4: Qualitative results. Our constructed functional 3D scene graph features a hierarchical structure and covers a more comprehensive range of manipulable objects.

Qualitative Comparison. Figure 4 provides a more intuitive comparison of different methods in real-world scenarios. OpenFunGraph and FunGraph often only construct a few tabletop-level objects and can hardly recognize their interactive units, let alone reliable functional edges. Regarding multi-level relationships, closely arranged similar carriers (*e.g.*, rows of drawers, adjacent cabinet doors) introduce strong ambiguity. OpenFunGraph and FunGraph often ignore the functional carriers altogether, instead hard-linking the interactive units directly to the object body; even so, under occlusion and viewpoint changes, this direct connection lacks geometric and topological grounding, and incorrect associations occur frequently. In contrast, our method can accurately restore typical hierarchical structures, and successfully construct most nodes and functional edges in tabletop scenarios. See the supplement for more qualitative results.

Real-World Manipulation Tasks. Furthermore, to validate the utility of the constructed hierarchical functional scene graph in real-world manipulation tasks, we conduct multi-task natural language-guided manipulation demonstrations on a physical robotic arm platform; experimental details and results are provided in the supplementary material.

6.3 Ablation Studies

Ablation Settings. We define three ablations on the FunGraph3D dataset [75] to evaluate the contribution of each key component: 1) *Assoc-OpenFG* replaces our multi-cued node association with OpenFunGraph’s 3D-IoU + CLIP cosine similarity; 2) *No-GO-Count* removes graph optimization, selecting the edge via Top-1 2D observation count. 3) *2D-Hierarchy* directly parses the hierarchical relationship at the 2D level without refinement at the 3D level;

Results and Analysis. As shown in Table 3, *Assoc-OpenFG* is significantly impaired: 3D-IoU has poor discriminability for adjacent similar instances (*e.g.*, bottles, drawers) and CLIP is unstable for parts, while our multi-cued scoring effectively suppresses such mismatches. *No-GO-Count* shows a decrease: simple counting cannot resolve 2D ambiguities, indicating frequency alone is insufficient to replace energy-based temporal graph optimization. *2D-Hierarchy* declines primarily in hierarchical triplets: parsing the $\mathcal{O} \leftarrow \mathcal{C} \leftarrow \mathcal{U}$ hierarchy in 2D introduces candidate explosion and view ambiguity, while moving this determination to 3D avoids these issues. See the supplement for more ablations.

7 Conclusion

This paper aims to address the deficiencies of existing functional 3D scene graphs in handling tabletop objects and hierarchical structures for fine-grained manipulation. We first extend the benchmark to cover tabletop manipulable objects and hierarchical structures. To tackle the challenges introduced by this expansion, which existing paradigms struggle to address, including the lack of visual anchoring in relational reasoning, instance confusion during cross-frame fusion, and attribution uncertainty under dynamic viewpoints, we propose an open-vocabulary, visually grounded and graph-optimization-based pipeline. This approach anchors functional edges using 2D evidence, stabilizes instance identity via multi-cue association in 3D, robustly determines edge attribution through temporal graph optimization, and finally performs global hierarchy shaping to recover the hierarchical structure. Experiments demonstrate that our method significantly outperforms existing approaches on the expanded benchmark.

Limitations. 1) Ambiguities may persist under extreme occlusion or restricted viewpoints; 2) Performance is bounded by the quality of detection/segmentation and VLM discrimination. Future work will explore active viewpoint selection.

Acknowledgements

This work was supported in part by the following funding sources: the National Science and Technology Major Project (Grant 2025ZD1606303); the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant JYB2025XDXM503); and the National Science Foundation of China (Grant 62495092). Additional support for this work was provided by the Swiss National Science Foundation Advanced Grant 216260, “Beyond Frozen Worlds: Capturing Functional 3D Digital Twins from the Real World”; and by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101214398 (ELLIOT), fully funded by the Swiss State Secretariat for Education, Research and Innovation (SERI). Alexandros Delitzas is also supported by a doctoral fellowship from the Max Planck ETH Center for Learning Systems (CLS).

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 preprint arXiv:2303.08774 (2023)
2. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: ICCV (2019)
3. Atzmon, M., Maron, H., Lipman, Y.: Point convolutional neural networks by extension operators. ACM TOG (2018)
4. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., et al.: Introducing hot3d: An ego-centric dataset for 3d hand and object tracking. arXiv preprint arXiv:2406.09598 (2024)
5. Bieri, V., Zamboni, M., Blumer, N.S., Chen, Q., Engelmann, F.: OpenCity3D: 3D Urban Scene Understanding with Vision-Language Models (2025)
6. Chen, L., Wang, X., Lu, J., Lin, S., Wang, C., He, G.: Clip-driven open-vocabulary 3d scene graph generation via cross-modality contrastive learning. In: CVPR (2024)
7. Chen, Z., Hasson, Y., Schmid, C., Laptev, I.: Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In: ECCV (2022)
8. Cho, W., Lee, J., Yi, M., Kim, M., Woo, T., Kim, D., Ha, T., Lee, H., Ryu, J.H., Woo, W., et al.: Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics. ECCV (2024)
9. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: CVPR (2019)
10. Delitzas, A., Parelli, M., Hars, N., Vlassis, G., Anagnostidis, S.K., Bachmann, G., Hofmann, T.: Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. In: BMVC (2023)
11. Delitzas, A., Takmaz, A., Tombari, F., Sumner, R., Pollefeys, M., Engelmann, F.: Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In: CVPR (2024)
12. Delitzas, A., Zhang, C., Gavryushin, A., Di Mario, T., Sun, B., Dabral, R., Guibas, L., Theobalt, C., Pollefeys, M., Engelmann, F., Barath, D.: Reconstructing Functional 3D Scenes from Egocentric Interaction Videos. In: CVPR (2026)
13. Do, T.T., Nguyen, A., Reid, I.: Affordancenet: An end-to-end deep learning approach for object affordance detection (2018)
14. Engelmann, F., Bokeloh, M., Fathi, A., Leibe, B., Nießner, M.: 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In: CVPR (2020)
15. Engelmann, F., Manhardt, F., Niemeyer, M., Tateno, K., Pollefeys, M., Tombari, F.: Opennerf: Open set 3d neural scene segmentation with pixel-wise features and rendered novel views. ICLR (2024)
16. Fan, Z., Parelli, M., Kadoglou, M.E., Chen, X., Kocabas, M., Black, M.J., Hilliges, O.: Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In: CVPR (2024)
17. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: CVPR (2018)
18. Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.S.: Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In: CVPR (2025)

19. 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 (2024)
20. Han, L., Zheng, T., Xu, L., Fang, L.: Occuseg: Occupancy-aware 3d instance segmentation. In: CVPR (2020)
21. Hou, J., Dai, A., Nießner, M.: 3d-sis: 3d semantic instance segmentation of rgb-d scans. In: CVPR (2019)
22. Hsu, J., Mao, J., Wu, J.: Ns3d: Neuro-symbolic grounding of 3d objects and relations. In: CVPR (2023)
23. Hu, X., Wu, Y., Zhao, M., Cao, Z., Zhang, X., Ji, X.: Dyo-slam: Visual localization and object mapping in dynamic scenes. *IEEE Transactions on Circuits and Systems for Video Technology* (2025)
24. Hu, Z., Bai, X., Shang, J., Zhang, R., Dong, J., Wang, X., Sun, G., Fu, H., Tai, C.L.: Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In: ICCV (2021)
25. Hua, B.S., Tran, M.K., Yeung, S.K.: Pointwise convolutional neural networks. In: CVPR (2018)
26. Huang, R., Peng, S., Takmaz, A., Tombari, F., Pollefeys, M., Song, S., Huang, G., Engelmann, F.: Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. *ECCV* (2024)
27. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3d visual grounding. In: CVPR (2022)
28. Huang, X., Huang, Y.J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., Zhang, L.: Open-set image tagging with multi-grained text supervision. In: *ACM MM* (2025)
29. Jatavallabhula, K.M., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Maalouf, A., Li, S., Iyer, G., Saryazdi, S., Keetha, N., et al.: Conceptfusion: Open-set multimodal 3d mapping. *ICRA2023 Workshop on Pretraining for Robotics (PT4R)* (2023)
30. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: Pointgroup: Dual-set point grouping for 3d instance segmentation. In: CVPR (2020)
31. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: ICCV (2023)
32. Koch, S., Hermosilla, P., Vaskevicius, N., Colosi, M., Ropinski, T.: Lang3dsg: Language-based contrastive pre-training for 3d scene graph prediction (2024)
33. 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: CVPR (2024)
34. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: CVPR (2018)
35. Li, Q., Mo, K., Yang, Y., Zhao, H., Guibas, L.: IFR-Explore: Learning inter-object functional relationships in 3d indoor scenes. *ICLR* (2022)
36. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. *NeurIPS* (2018)
37. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
38. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (2024)
39. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: *NeurIPS* (2023)
40. Nagaraajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: ICCV (2019)

41. Ok, K., Liu, K., Frey, K., How, J.P., Roy, N.: Robust object-based slam for high-speed autonomous navigation. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 669–675. IEEE (2019)
42. OpenAI: Gpt-5 system card. Technical report, OpenAI (Aug 2025), <https://cdn.openai.com/gpt-5-system-card.pdf>, accessed: 2025-11-14
43. Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., Hofmann, T.: CLIP-Guided Vision-Language Pre-Training for Question Answering in 3D Scenes. In: CVPRW (2023)
44. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: CVPR (2023)
45. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR (2017)
46. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. NeurIPS (2017)
47. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: CVPR (2024)
48. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024)
49. Roh, J., Desingh, K., Farhadi, A., Fox, D.: Languagerefer: Spatial-language model for 3d visual grounding (2022)
50. Rosinol, A., Gupta, A., Abate, M., Shi, J., Carlone, L.: 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. Robotics, Science and Systems (2020)
51. 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 (2021)
52. Rotondi, D., Scaparro, F., Blum, H., Arras, K.O.: Fungraph: Functionality aware 3d scene graphs for language-prompted scene interaction (2025)
53. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation (2023)
54. Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., Yan, Z.: Going denser with open-vocabulary part segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
55. Takmaz, A., Delitzas, A., Sumner, R.W., Engelmann, F., Wald, J., Tombari, F.: Search3D: Hierarchical Open-Vocabulary 3D Segmentation (2025)
56. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. NeurIPS (2023)
57. Takmaz, A., Schult, J., Kaftan, I., Akçay, M., Leibe, B., Sumner, R., Engelmann, F., Tang, S.: 3D Segmentation of Humans in Point Clouds with Synthetic Data. In: ICCV (2023)
58. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: ICCV (2019)
59. Vu, T., Kim, K., Luu, T.M., Nguyen, T., Yoo, C.D.: Softgroup for 3d instance segmentation on point clouds. In: CVPR (2022)
60. Wald, J., Dhano, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: CVPR (2020)
61. Wang, Z., Cheng, B., Zhao, L., Xu, D., Tang, Y., Sheng, L.: Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In: CVPR (2023)

62. Weder, S., Blum, H., Engelmann, F., Pollefeys, M.: Labelmaker: Automatic semantic label generation from rgb-d trajectories (2024)
63. Weder, S., Engelmann, F., Schönberger, J.L., Seki, A., Pollefeys, M., Oswald, M.R.: Alster: A Local Spatio-temporal Expert for Online 3D Semantic Reconstruction (2023)
64. Wei, M., Yue, X., Zhang, W., Kong, S., Liu, X., Pang, J.: Ov-parts: Towards open-vocabulary part segmentation. *Advances in Neural Information Processing Systems* **36**, 70094–70114 (2023)
65. Wu, S.C., Tateno, K., Navab, N., Tombari, F.: Incremental 3d semantic scene graph prediction from rgb sequences. In: *CVPR* (2023)
66. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: SceneGraphFusion: Incremental 3d scene graph prediction from rgb-d sequences. In: *CVPR* (2021)
67. Yang, Z., Zhang, S., Wang, L., Luo, J.: Sat: 2d semantics assisted training for 3d visual grounding. In: *ICCV* (2021)
68. Ye, Y., Gupta, A., Kitani, K., Tulsiani, S.: G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In: *CVPR* (2024)
69. Ye, Y., Gupta, A., Tulsiani, S.: What’s in your hands? 3d reconstruction of generic objects in hands. In: *CVPR* (2022)
70. Yilmaz, G., Peng, S., Pollefeys, M., Engelmann, F., Blum, H.: OpenDAS: Open-Vocabulary Domain Adaptation for 2D and 3D Segmentation. *arXiv preprint arXiv:2405.20141* (2024)
71. Yoshida, T., Kurita, S., Nishimura, T., Mori, S.: Text-driven affordance learning from egocentric vision. *arXiv preprint arXiv:2404.02523* (2024)
72. Yue, Y., Mahadevan, S., Schult, J., Engelmann, F., Leibe, B., Schindler, K., Kontogianni, T.: Agile3d: Attention guided interactive multi-object 3d segmentation. *ICLR* (2024)
73. Zhai, W., Luo, H., Zhang, J., Cao, Y., Tao, D.: One-shot object affordance detection in the wild. *IJCV* (2022)
74. Zhang, C., Yu, J., Song, Y., Cai, W.: Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In: *CVPR* (2021)
75. 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: *CVPR* (2025)
76. Zhang, C., Di, Y., Zhang, R., Zhai, G., Manhardt, F., Tombari, F., Ji, X.: Ddf-ho: Hand-held object reconstruction via conditional directed distance field. *NeurIPS* (2023)
77. Zhang, C., Jiao, G., Di, Y., Wang, G., Huang, Z., Zhang, R., Manhardt, F., Fu, B., Tombari, F., Ji, X.: Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In: *CVPR* (2024)
78. Zhang, S., Hao, A., Qin, H., et al.: Knowledge-inspired 3d scene graph prediction in point cloud. *NeurIPS* (2021)
79. Zhang, Y., Gong, Z., Chang, A.X.: Multi3drefer: Grounding text description to multiple 3d objects. In: *ICCV* (2023)
80. 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: *CVPR* (2024)
81. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: *CVPR* (2024)
82. Zuo, X., Samangouei, P., Zhou, Y., Di, Y., Li, M.: Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *IJCV* (2024)