Beyond Isolated Objects: Relationship-aware Open Vocabulary Scene Understanding via 3D Scene Graph Analysis

Xianhao Chen[1*], Jiarui Hu[1*], Yuanbo Yang[2], Xiyu Zhang[1], Tengyue Wang[2], Hujun Bao[1], Guofeng Zhang[1], and Zhaopeng Cui[1†]

[1] State Key Lab of CAD&CG, Zhejiang University, China
[2] Zhejiang University, China

Abstract. Open-vocabulary 3D scene understanding aims to segment 3D scenes beyond predefined categories by transferring semantic knowledge from vision-language models. Existing methods have advanced this task by lifting language-aligned 2D features into 3D, yet they often rely on context-independent semantic representations, leaving object relationships underexplored for contextual refinement. We propose *RelGraphOV*, a relationship-aware framework that uses 3D scene graphs to enhance open-vocabulary 3D understanding. Our method constructs relational scene graphs from multi-view observations by leveraging vision-language reasoning to infer object relationships and prune geometrically implausible connections, without manual relationship annotations. To aggregate relational context while avoiding feature interference, we introduce an Adaptive Gated Dual-Stream Contextual GAT that separates dense geometric features and semantic CLIP embeddings, performs edge-guided message passing, and adaptively fuses complementary semantics. A hierarchical contrastive objective further promotes instance-level consistency and category-level discrimination. Experiments on ScanNetV2, ScanNet200, ScanNet++, and Replica demonstrate strong performance and generalization ability. Project Page: cxavireh.github.io/relgraphov-projectpage

Keywords: 3D Semantic Segmentation · Scene Graph · Open-Vocabulary

1 Introduction

Open-vocabulary 3D scene understanding primarily aims to perform dense semantic segmentation of 3D environments, extending object recognition beyond a fixed set of predefined categories to enable scalable perception for applications such as robotic navigation, spatial reasoning, and augmented reality. Inspired by vision-language models such as CLIP [35], recent works transfer semantic knowledge learned from image-text data into 3D representations. Existing open-vocabulary 3D semantic segmentation approaches generally fall into two

* Equal contribution. † Corresponding author.

Fig. 1: RelGraphOV. We propose a relationship-aware open-vocabulary 3D scene understanding framework. Left: Illustration of encoding a 3D scene into a relationship-aware scene graph to capture contextual relationships. Right: Qualitative open-vocabulary segmentation results produced by our method.

paradigms. The first projects *dense vision-language features* from 2D models into 3D space [25, 33], which provides strong spatial robustness under occlusion or incomplete observations. The second associates *instance-level CLIP embeddings* with 3D object proposals [40, 44, 48], enabling stronger semantic generalization to long-tail categories. Despite their complementary strengths, both paradigms primarily rely on context-independent semantic representations, leaving structured object relationships underexplored for semantic refinement. For instance, distinguishing a "curtain" from a "shower curtain" is notoriously difficult when relying exclusively on isolated object appearances. However, if the model is aware of the surrounding context, such as the object's spatial relationship to a bathtub or a toilet, this semantic ambiguity can be naturally resolved.

A natural representation for modeling such contextual information is the *3D scene graph*, which represents objects as nodes and their relationships as edges. Existing relational 3D methods have studied scene graph prediction [19, 20, 42, 45, 47], relationship querying [12, 21], referred object grounding [3], and relation-aware instance segmentation [29]. However, they do not directly address how inferred 3D semantic relationships can serve as intermediate contextual priors to refine open-vocabulary semantic features for dense 3D segmentation. Introducing scene graphs into this setting poses two major challenges. First, *data preparation* becomes difficult because open-vocabulary settings significantly increase the diversity of objects and relationships, while large-scale relationship annotations are rarely available. Second, aggregating heterogeneous representations on the graph leads to a fundamental *feature granularity dilemma*. Dense features provide strong spatial robustness but limited semantic generalization, whereas CLIP-based instance features offer richer semantics but are sensitive to partial observations. Although these representations appear complementary, naively combining them within graph neural networks often causes severe feature interference, where geometric noise contaminates the semantic embedding space during message passing.

To address these challenges, we propose *RelGraphOV* (see Fig. 1), a relationship aware open-vocabulary 3D scene understanding framework that explicitly incorporates contextual reasoning via 3D scene graphs. Our framework first constructs a relational scene graph from multi-view observations and then performs contextual feature learning on the graph to enhance open-vocabulary representations. Specifically, to alleviate the lack of relationship annotations, we design a scene graph construction pipeline that leverages multi-view observations and vision-language reasoning to infer object relationships and filter geometrically implausible connections. This process enables labor-free and efficient graph construction without manual relationship labeling and provides reliable relational priors for contextual feature aggregation. Different from prior relational 3D methods that mainly target scene graph construction, querying, grounding, or instance-level reasoning, our approach uses the constructed graph as a relational structure for contextual feature learning in open-vocabulary 3D segmentation.

To further address the feature granularity dilemma and prevent feature interference during graph-based aggregation, we propose an Adaptive Gated Dual-Stream Contextual GAT. Instead of directly mixing heterogeneous representations, our architecture decouples feature propagation into two streams. The main stream aggregates contextual information using dense LSeg features along scene graph edges, while the auxiliary stream independently preserves high-level semantic embeddings extracted from CLIP. A learnable gating mechanism dynamically injects complementary semantic cues from the auxiliary stream during message passing, enriching node representations while preventing cross-modal contamination. In addition, we introduce a hierarchical contrastive learning strategy that jointly enforces instance-level consistency and category-level discrimination, effectively mitigating semantic drift in open-vocabulary settings.

We evaluate our method on ScanNetV2 [7], ScanNet200 [37], and further conduct strict cross dataset zero-shot evaluation on ScanNet++ [51] and Replica [39]. Experimental results demonstrate that *RelGraphOV* consistently outperforms existing methods and achieves state-of-the-art performance in open-vocabulary 3D scene understanding.

To summarize, our contributions are as follows:

- We propose *RelGraphOV*, a relationship-aware open-vocabulary 3D scene understanding framework that incorporates contextual reasoning via automatically constructed 3D scene graphs.
- We introduce a multi-view reasoning-based scene graph construction pipeline that enables scalable generation of object relationships without manual annotation.
- We propose an Adaptive Gated Dual-Stream Contextual GAT with hierarchical contrastive learning to address the feature granularity dilemma and mitigate feature interference during graph-based context aggregation.
- Extensive experiments on ScanNetV2, ScanNet200, ScanNet++, and Replica demonstrate state-of-the-art performance and strong zero-shot generalization.

2 Related Work

2D Visual Foundation Models. Recent advances in large-scale pre-training have established Visual Foundation Models (VFMs) as the cornerstone for 2D visual perception. CLIP [35] aligns images and text through contrastive pre-training, enabling zero-shot recognition, while DINO [2, 32] learn transferable dense visual features. These models have been widely used for segmentation [4, 16, 18, 36, 54], detection [28], captioning [27, 52], and open-vocabulary recognition [5, 30]. In 3D scene understanding, they serve as semantic priors for transferring language-aligned knowledge from 2D observations to 3D representations.

Open-Vocabulary 3D Scene Understanding. Conventional 3D semantic segmentation methods [6,31] are typically trained within predefined label spaces and require dense annotations, limiting their generalization to novel categories. Recent open-vocabulary 3D methods address this limitation by transferring 2D vision-language features into 3D scenes. OpenScene [33] pioneered this direction by distilling dense language-aligned 2D features, such as LSeg [24] and OpenSeg [11], into 3D point clouds. Subsequent methods [10,17,23,25,26,50] improve 2D–3D feature alignment, multi-view feature fusion, and open-vocabulary querying. Instance-level methods such as OpenMask3D [40] and related approaches [44] further associate object proposals with CLIP-like features for object-level recognition. While effective, these methods often treat objects in a scene as isolated entities, leaving structured object relations underexplored for context-aware semantic refinement.

Relational 3D Scene Understanding. Object relations have been explored in 3D scene graph prediction and relation-aware perception. Existing 3D scene graph methods predict object nodes and semantic relationships from reconstructed scenes or RGB-D sequences [12, 42, 47], with recent works further incorporating visual-linguistic supervision, language-based pre-training, or open-vocabulary relationship prediction [19, 20, 45]. Beyond scene graph prediction, relations have also been used for referred entity grounding [3], relational querying in radiance fields [21], and closed-set instance segmentation via superpoint-level geometric relations [29]. These studies show the value of explicit object relations for 3D scene understanding. Our work studies a complementary use of such relations in open-vocabulary 3D semantic segmentation: the constructed scene graph is used as a contextual structure for aggregating features among object proposals, so that language-aligned 3D features can be refined before dense point-level labeling.

3 Method

Problem Statement. As illustrated in Fig. 2, 3D open-vocabulary scene understanding aims to segment 3D scenes with free-form textual descriptions, going beyond a fixed set of predefined categories while supporting natural-language queries. The system processes each point in the scene by aligning its 3D features

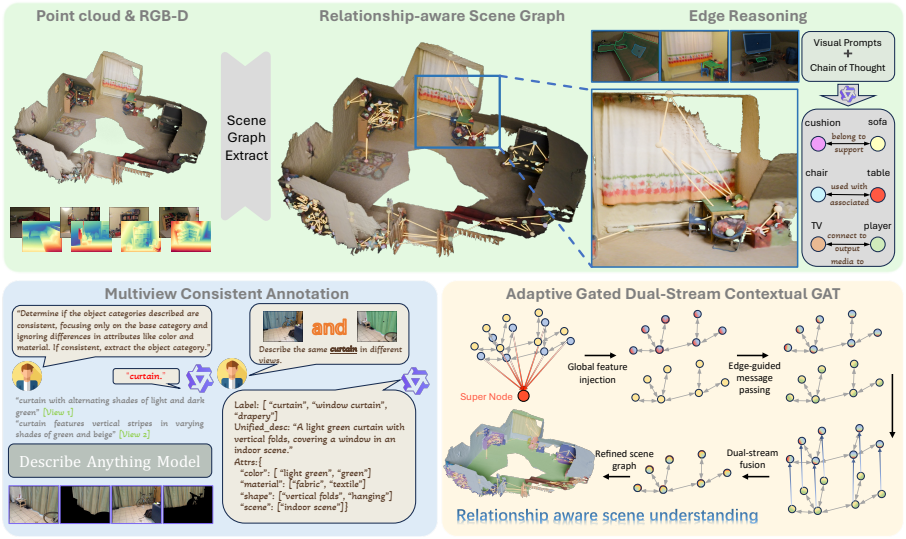

Fig. 2: Overall Pipeline of RelGraphOV. (1) *Graph Construction*: We build a relationship-aware 3D scene graph via multi-view VLM reasoning. (2) *Annotation*: A VLM chain-of-thought generates rich node semantics. (3) *Adaptive Gated Dual-Stream GAT*: To prevent feature interference, we decouple geometric (main) and semantic (auxiliary) feature propagation. Both undergo edge-guided message passing, while a learnable gate dynamically injects auxiliary semantics into the main stream. (4) *Training*: Hierarchical contrastive and dual-alignment losses optimize the network, effectively mitigating semantic drift and catastrophic forgetting.

with semantic embeddings derived from natural language. This enables accurate point-level labeling and retrieval of semantically relevant regions based on textual descriptions.

Overview. As illustrated in Fig. 2, the input to our system consists of RGB-D sequences and point clouds, where the point clouds can be obtained either from SLAM algorithms [1, 13, 14] or dense 3D reconstruction [43, 53]. The first step (Sec. 3.1) is to abstract the scene into a relationship-aware 3D scene graph, where keyframes are extracted using a unified viewpoint scoring mechanism, node features are initialized, and semantic relationships for each edge are reasoned by a Vision-Language Model (VLM). In the second step (Sec. 3.2), we generate node annotations, embedding rich semantic information by leveraging the powerful captioning capabilities of a VLM. The third step (Sec. 3.3) involves employing an Adaptive Gated Dual-Stream Contextual GAT, which decouples multi-modal feature propagation and utilizes edge-feature-guided message passing to optimize node semantics without suffering from feature interference. Finally, we design a hierarchical contrastive loss with auxiliary supervision (Sec. 3.4) to train the model based on its dual-stream characteristics and the 3D scene graph structure.

3.1 Relationship-Aware Scene Graph Construction

To effectively utilize the environmental context within the scene for improved scene understanding, it is crucial to construct a relationship-aware open vocabulary 3D scene graph. First, following MaskClustering [48], we aggregate 2D masks from RGB-D sequences to obtain 3D class-agnostic object proposals, represented as point cloud clusters $\mathcal{M} = \{M_1, M_2, \ldots, M_N\}$. We treat such proposal generation and feature extraction as interchangeable upstream modules rather than our focus, so RelGraphOV inherits their proposal quality while remaining agnostic to their specific choice. This modular design also allows our relational refinement to directly benefit from increasingly powerful proposal generators and vision-language backbones without architectural changes. Taking these 3D proposals as input, the scene graph construction process comprises four key steps.

Node Definition, Keyframe Extraction, and Pruning. We treat each proposal M_i as a candidate node v_i and filter out small, meaningless instances based on the number of points. For each node v_i, we calculate its 3D axis-aligned bounding box B_i and centroid C_i. At this stage, to obtain comprehensive visual observations for each node, we extract representative keyframes from the original RGB-D sequence. Specifically, by projecting the 3D bounding box onto the 2D frames, we evaluate each candidate frame I_k using a universal composite scoring function $S(I_k, W)$:

$$\begin{aligned} S(I_k, W) = W \cdot \mathrm{S}, \quad W = \left[w_{\text{comp}}, w_{\text{area}}, w_{\text{center}}, w_{\text{sharp}}\right], \\ \mathrm{S} = \left[s_{\text{comp}}, s_{\text{area}}, s_{\text{center}}, s_{\text{sharp}}\right]^T, \end{aligned} \tag{1}$$

where $s_{\text{comp}}, s_{\text{area}}, s_{\text{center}}, s_{\text{sharp}} \in (0, 1)$ represent the node completeness, projected area, center score, and motion blur sharpness, respectively. W is a customizable weight vector tailored to different selection objectives. Using this mechanism, we select keyframes offering optimal visibility and centrality for each object. Following this, we perform structural pruning. Nodes representing large, non-interactive background structures, such as floors, walls, and ceilings, are identified and removed based on specific geometric attributes, including volume and surface normal direction. This ensures the graph focuses exclusively on core interactive objects and avoids redundant connections.

Node Feature Initialization. With the keyframes extracted, we proceed to acquire the multi-modal features for each node. We project the 3D point cloud of each node onto its corresponding 2D keyframes to extract pixel-aligned dense features using LSeg [24] and instance-level global features using CLIP [35]. The extracted 2D features are pooled to form the base 3D LSeg semantic feature f_l and instance semantic CLIP feature f_c for each node. Consequently, the comprehensive multimodal feature representation for each node is formulated as $\mathbf{F}_v = \{f_l, f_c\}$, which serves as the decoupled inputs for our dual-stream architecture.

Edge Initialization. We initialize the graph's connectivity $\mathbf{E}_{geom}$ based on spatial proximity. An edge e_{ij} is established either if the nodes are among the three nearest neighbors based on Euclidean distance between centroids, or if

their 3D bounding boxes are adjacent or overlap. This constructs a dense and geometrically coherent candidate graph.

VLM-based Edge Refinement and Encoding. The final step is the core of our relational construction, where we refine each candidate edge $e_{ij} \in \mathbf{E}_{geom}$ using a VLM. Specifically, we first select a representative frame $I_{v_i v_j}$ that clearly captures both nodes by jointly applying the unified scoring criterion (Eq. 1) to their projected regions. Given this optimal viewpoint, we draw inspiration from Set-of-Mark [49] to visually prompt the VLM agent, employing a chain-of-thought [46] reasoning process to explicitly describe the inter-object relations. Based on the VLM's response, any candidate edge returning a "*none*" relationship is immediately pruned. The subset of spatial connections successfully validated by the VLM constitutes our refined edge set, denoted as $\mathbf{E}_{refine}$. For these valid edges, the VLM generates two asymmetric descriptions (e.g., "*person sits on chair*" and "*chair supports person*"). To formally encode the edge feature $\mathbf{F}_e$, we parse these descriptions to extract Subject-Verb-Object (SVO) triplets. We then feed the subject, verb, and object independently into the CLIP text encoder and concatenate the resulting three feature vectors to construct a highly-structured edge representation. This process yields the refined relationship-aware scene graph:

$$\mathbf{SG} = (\mathbf{V}, \mathbf{E}_{refine}, \mathbf{F}_v, \mathbf{F}_e). \tag{2}$$

3.2 VLM-based Scene Graph Annotation Engine

The constructed graph requires accurate node-level annotations for supervision. However, the same 3D object viewed from different 2D perspectives may yield conflicting descriptions. We design an automated engine to adjudicate this multi-view inconsistency.

Multi-View Selection and View-Specific Captioning. For each node v_i, we project its point cloud onto all RGB-D frames. To rank these frames, we utilize the unified composite scoring function $S(I_k, W)$ defined in Eq. 1. By adjusting the weight vector W, we first choose the best viewpoint I_1 emphasizing instance visibility. Then, we select a second viewpoint I_2 by applying a different set of weights emphasizing centrality and sharpness, accompanied by an IoU penalty to ensure spatial dissimilarity. These viewpoints are fed into SAM [18] with sampled point prompts to generate object masks M_1^{2D} and M_2^{2D}, which are subsequently processed by the Describe Anything Model [27] to generate initial viewpoint-specific descriptions D_1 and D_2.

VLM Adjudication and Refinement. To ensure consistency, a lightweight LLM assesses if the core category recognition in D_1 and D_2 aligns. If consistent, a guided prompt instructs the VLM to refine attributes using both views. Otherwise, an adjudication prompt explicitly informs the VLM of the conflict, directing it to use visual content from both views as the "final truth" to determine the correct category. This effectively filters out background noise and yields highly accurate textual annotations. The final textual annotation of node v_i is encoded by the CLIP text encoder as f_{anno}^i, which is used as soft semantic supervision during training.

3.3 Adaptive Gated Dual-Stream Contextual GAT

After obtaining the scene graph **SG** and VLM edge features $\mathbf{F}_e$, our objective is to refine the context-independent features into context-aware representations. However, naively mixing LSeg features f_l with instance-level CLIP features f_c during graph aggregation causes severe feature interference. To resolve this, we propose an Adaptive Gated Dual-Stream GAT. To ensure mathematical clarity, we use subscript m to denote the main stream and a to denote the auxiliary stream. The network architecture and refinement process consist of the following carefully decoupled modules.

Modality-Specific Node Initialization. We establish two decoupled initial states to prevent early contamination. The center point coordinates and bounding box dimensions are encoded via a Fourier Feature Encoder and an MLP. These geometric features are concatenated with the dense LSeg features (f_l) to form the initial state $h_{m,i}^{(0)}$ for the main stream, and concatenated with the projected CLIP features (f_c) to form the initial state $h_{a,i}^{(0)}$ for the auxiliary stream.

Global Context Aggregation. To prevent over-smoothing and expand the receptive field, we dynamically introduce a virtual ***super node*** v_{super} connected to all real nodes v_i, forming an augmented graph G_{aug}. Processed by a standard GAT module, the output state of v_{super} serves as the global context feature G_{global}.

Global Context Injection. During each layer l of the GAT, we first inject the global context exclusively into the main stream via a Cross-Attention module. Using the main stream node state $h_{m,i}^{(l)}$ as the Query, and the global feature G_{global} as both Key and Value, the context-modulated feature is merged with $h_{m,i}^{(l)}$ through an MLP. This produces a global-aware state $\tilde{h}_{m,i}^{(l)}$, explicitly bridging local geometry with scene-level understanding.

Edge-Feature-Guided Message Propagation. Following the global injection, both streams perform independent neighborhood aggregation. To emphasize the critical role of relationship semantics, we design an edge-feature-guided propagation mechanism. Specifically, for the main stream, we use the node's global-aware state $\tilde{h}_{m,i}^{(l)}$ as the **Query**. For the **Key**, we concatenate the neighbor node's state $\tilde{h}_{m,j}^{(l)}$ with the edge feature $f_{e_{ij}}$. The **Value** is strictly defined as the neighbor node's state $\tilde{h}_{m,j}^{(l)}$. By injecting $f_{e_{ij}}$ into the Key, the cross-attention mechanism adaptively modulates the attention weights, forcing the GNN to amplify information from strongly relevant neighbors and suppress irrelevant ones based solely on their semantic relationships. The auxiliary stream undergoes a similarly independent propagation using its own state $h_{a,i}^{(l)}$. We denote the intermediate aggregated states after this step as $\hat{h}_{m,i}^{(l)}$ and $\hat{h}_{a,i}^{(l)}$.

Adaptive Gated Fusion. After the independent message passing, the final step in the layer update is to allow the main stream to dynamically borrow richer, highly generalizable semantics from the auxiliary stream. We achieve this via a context-aware fusion gate:

$$Gate = \sigma(\text{MLP}([\hat{h}_{m,i}^{(l)} \parallel \hat{h}_{a,i}^{(l)}])), \tag{3}$$

$$h_{m,i}^{(l+1)} = \hat{h}_{m,i}^{(l)} + \alpha \cdot Gate \odot \hat{h}_{a,i}^{(l)}, \tag{4}$$

where α is a scaling factor. Concurrently, the auxiliary stream securely preserves its pure semantics for the next layer ($h_{a,i}^{(l+1)} = \hat{h}_{a,i}^{(l)}$). This unilateral gating mechanism allows autonomous calibration of the main semantic representation without contaminating the auxiliary open-vocabulary priors.

Feature Decoders. After L layers, the main stream feature $h_{m,i}^{(L)}$ is processed by an MLP decoder to yield the final context-aware prediction F_{out}. Concurrently, the auxiliary stream employs an independent decoder to output F_{aux}, used exclusively for auxiliary supervision.

3.4 Dual-Stream Model Training

Learning Objective. The training objective of our model is to learn a refinement process that transforms context-independent features F_v into context-aware features F_{out}. Since we use the open-vocabulary annotations generated in Sec. 3.2 as supervision, a standard contrastive loss would penalize semantically similar categories. We design a Hierarchical Contrastive Loss ($\mathcal{L}_{hie}$) for the main stream to solve this issue, along with dual-alignment losses ($\mathcal{L}_{reg}$ and $\mathcal{L}_{aux}$) to prevent catastrophic forgetting.

Loss Design. We design a composite loss function:

$$\mathcal{L}_{total} = \lambda_{\text{hie}} \cdot \mathcal{L}_{\text{hie}} + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}} + \lambda_{\text{aux}} \cdot \mathcal{L}_{\text{aux}}. \tag{5}$$

The core term is $\mathcal{L}_{\text{hie}}$, which encourages the refined node feature F_{out} to satisfy three objectives: (1) self-consistency with its corresponding annotation feature f_{anno}^{i}, (2) intra-class consistency among semantically similar instances, and (3) inter-class separation from unrelated categories.

To model intra-class consistency, we construct an inferred positive mask M_{pos}, where $M_{pos}^{ij} = 1$ if the cosine similarity between annotation features f_{anno}^{i} and f_{anno}^{j} exceeds a threshold τ. To further emphasize self-consistency, we introduce a margin ($\mathcal{M}$) to the self-pair similarity:

$$S_{ij} = \frac{F_{out}^{v_i} \cdot f_{anno}^{j}}{\eta} + \delta_{ij} \cdot \mathcal{M}, \tag{6}$$

$$\mathcal{L}_{hie} = -\frac{1}{|\mathbf{V}|} \sum_{i \in \mathbf{V}} \sum_{j \in \mathbf{V}} \log \frac{M_{\text{pos}}^{ij} \exp(S_{ij})}{\sum_{k \in \mathbf{V}} \exp(S_{ik})}, \tag{7}$$

where η denotes the temperature parameter and δ_{ij} is the Kronecker delta.

In addition, both the regularization loss $\mathcal{L}_{reg}$ and the auxiliary loss $\mathcal{L}_{aux}$ follow a shared dual-alignment formulation designed to balance learning new open-vocabulary semantics while preserving modality-specific priors. We define a generalized alignment loss:

$$\mathcal{L}_{align}(F, f_{init}, \gamma) = \frac{1}{N} \sum_{i=1}^{N} \left[\gamma(1 - F^{v_i} \cdot f_{anno}^{i}) + (1-\gamma)(1 - F^{v_i} \cdot f_{init}^{v_i}) \right], \tag{8}$$

where f_{init} denotes the initial modality feature and γ controls the trade-off between adapting to VLM annotations and preserving the original feature prior.

Accordingly, we define $\mathcal{L}_{reg} = \mathcal{L}_{align}(F_{out}, f_l, \gamma_{reg})$ to regularize the main stream by aligning it with both the VLM annotation f_{anno} and the dense feature prior f_l. Similarly, $\mathcal{L}_{aux} = \mathcal{L}_{align}(F_{aux}, f_c, \gamma_{aux})$ guides the auxiliary stream to learn from the same annotations while remaining anchored to its original CLIP representation f_c. This symmetric dual-alignment design enables both streams to absorb contextual supervision while preserving their complementary modality strengths during graph propagation.

4 Experiments

4.1 Experimental Setup

Datasets. To evaluate the effectiveness of our proposed method, we conduct experiments on widely used indoor 3D datasets, ScanNetV2 [7] and ScanNet200 [37], along with ScanNet++ [51] and the photorealistic synthetic dataset Replica [39]. For ScanNetV2 [7], we follow the official data split (1,201 scenes for training, 312 for validation) and adhere to the standard 20-class evaluation subset to ensure a fair comparison with prior work. ScanNet200 [37] extends the original annotations to a 200-category label space, offering a more fine-grained and challenging benchmark to assess model robustness in open-vocabulary scenarios. We further evaluate cross-dataset zero-shot generalization on ScanNet++ and Replica without additional training or fine-tuning, following the standard 100-class benchmark for ScanNet++ and the 51-class evaluation protocol for Replica.

Implementation Details. We run our system on a desktop equipped with an Intel i9-14900K CPU and an NVIDIA RTX 4090 GPU. Edge reasoning and node annotation generation are performed using qwen-vl-max [34], while qwen-turbo [34] is employed to verify category consistency as described in Sec. 3.2. During training, we utilize the Adam optimizer with an initial learning rate of 1×10^{-3}, training the model for 600 epochs. The batch size is set to 256 for all datasets. The loss weights $[\lambda_{\text{hie}}, \lambda_{\text{reg}}, \lambda_{\text{aux}}]$ are set to $[0.5, 1.5, 1.5]$, and the dual-alignment balance factors $[\gamma_{\text{reg}}, \gamma_{\text{aux}}]$ are set to $[0.7, 0.7]$, based on preliminary hyperparameter tuning. Further implementation specifics, including the exact VLM prompts, are provided in the supplementary material.

Metrics. We select open-vocabulary semantic segmentation as a representative task for evaluating our model's open-vocabulary scene understanding capabilities. Following the experimental setup in OpenScene [33], we use mIoU (%) and mAcc (%) as the evaluation metrics.

4.2 Open-Vocabulary Scene Understanding

As shown in Tab. 1, RelGraphOV achieves the best open-vocabulary mIoU on ScanNetV2 [7], competitive mAcc, and the best mIoU and mAcc on ScanNet200 [37] among the compared methods. These results suggest that relational context

Table 1: Open-Vocabulary 3D Semantic Segmentation Results on ScanNetV2 and ScanNet200. We report mIoU and mAcc. † denotes numbers reported in the original paper (ScanNetV2) or reproduced by us (ScanNet200) based on LSeg.

Type	Method	ScanNetV2		ScanNet200	
		mIoU	mAcc	mIoU	mAcc
Closed Vocabulary	TangentConv [41]	40.9	–	–	–
	TextureNet [15]	54.8	–	–	–
	ScanComplete [8]	56.6	–	–	–
	DCM-Net [38]	65.8	–	–	–
	Mix3D [31]	73.6	–	–	–
	MinkowskiNet [6]	69.2	77.7	–	–
Open Vocabulary	MSeg Voting [22]	45.6	54.4	–	–
	PLA [9]	17.7	33.5	1.8	3.1
	RegionPLC [50]	43.8	65.6	6.5	15.9
	OpenMask3D [40]	34.0	–	10.5	–
	MaskClustering [48]	34.4	59.6	11.8	24.7
	OpenScene-2D† [33]	50.0	62.7	8.8	13.5
	OpenScene-3D† [33]	52.9	63.2	6.8	8.7
	OpenScene-2D3D† [33]	54.2	66.6	7.9	11.4
	DMA-text only [26]	50.5	63.7	8.7	11.3
	Mosaic3D [23]	48.9	**73.6**	12.4	25.1
	CUA-O3D [25]	55.3	65.6	6.4	8.8
	OV3D [17]	57.3	72.9	8.7	–
	RelGraphOV (Ours)	**58.4**	73.4	**14.5**	**26.1**

Table 2: Per-Category IoU on ScanNetV2 [7]. We report IoU (%) for a representative subset of categories.

	cabinet	bed	door	window	fridge	curtain	shower curtain	toilet	bathtub	chair
OpenScene [33]	43.6	71.8	49.8	48.2	49.4	53.2	0.0	67.3	61.4	57.9
CUA-O3D [25]	44.7	73.2	50.5	52.8	39.6	58.7	0.0	78.4	62.3	**76.4**
Ours	**46.3**	**76.8**	**54.5**	**58.3**	**57.8**	**66.6**	**64.2**	**83.6**	**67.5**	71.7

and the Adaptive Gated Dual-Stream Contextual GAT improve open-vocabulary feature refinement across both standard and long-tail label spaces.

On ScanNetV2 [7], our method improves over OV3D [17] in mIoU and over OpenScene-2D [33] by +8.4% mIoU and +10.7% mAcc. The distinction between "shower curtain" and the general "curtain" category provides a representative case, as shown in Tab. 2 and Fig. 3. OpenScene [33] and CUA-O3D [25] fail to recognize "shower curtain" in this setting, likely because they rely mainly on object appearance and group the points into the broader "curtain" class. In contrast, RelGraphOV uses VLM-reasoned neighborhood relations during message passing, improving the IoU for "shower curtain" to 64.2%.

We further evaluate the open-vocabulary capability and generalization performance of RelGraphOV on the challenging ScanNet200 [37] benchmark. This dataset contains 200 categories, ten times larger than commonly used previous benchmarks, posing severe long-tail challenges. Existing methods struggle here

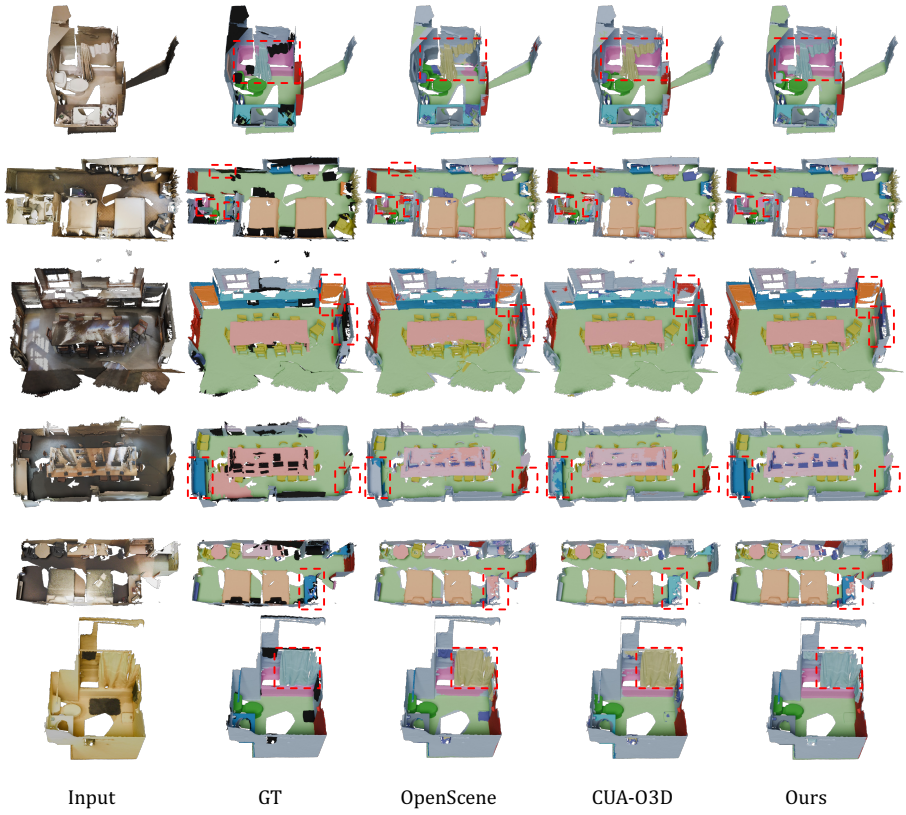

Fig. 3: Qualitative Comparisons. Images of 3D semantic segmentation results on ScanNetV2 [7], including the example of a curtain and shower curtain as mentioned in the introduction.

for two distinct reasons. Distillation-based approaches (e.g., OpenScene [33], DMA [26]) are constrained by the semantic coverage of their 2D teacher models. Conversely, methods directly using CLIP to extract instance features (e.g., MaskClustering [48], OpenMask3D [40]) have an advantage on long-tail categories but often struggle on large, texture-less base objects like walls and floors due to incomplete multi-view geometric observations. On ScanNet200, RelGraphOV obtains 14.5% mIoU and 26.1% mAcc, outperforming the compared open-vocabulary baselines in both metrics. This result suggests that the dual-stream design helps combine the spatial robustness of dense LSeg features with the category generalization of CLIP features.

To further analyze the source of the improvement on ScanNet200, we split its categories into head, common, and tail groups. As shown in Tab. 3, *RelGraphOV* achieves the best overall performance and shows clear advantages on common and tail categories. Although Mosaic3D performs better on head cat-

Table 3: Analysis on ScanNet200 [37]. Following the official split, the 200 categories are divided into head, common, and tail groups. Each entry reports mIoU/mAcc (%).

Method	All	Head	Common	Tail
OpenScene-2D [33]	8.8/13.5	19.6/29.2	3.7/8.4	1.5/2.4
MaskClustering [48]	11.8/24.7	20.7/34.3	8.6/20.6	5.5/17.5
Mosaic3D [23]	12.4/25.1	**28.9/43.1**	5.6/20.5	1.3/9.8
CUA-O3D [25]	6.4/8.8	19.0/26.3	0.2/0.6	0.0/0.0
RelGraphOV	**14.5/26.1**	26.1/36.3	**10.2/22.6**	**6.3/18.1**

Table 4: Zero-Shot Evaluation on ScanNet++ [51]. All methods are trained only on ScanNetV2 and evaluated directly on the 100-class ScanNet++ benchmark without fine-tuning.

Method	OpenScene-2D [33]	OpenScene-3D [33]	OpenScene-2D3D [33]	CUA-O3D [25]	**RelGraphOV**
mIoU	13.3	8.9	11.7	9.3	**20.9**
mAcc	20.0	13.6	15.3	13.7	**33.2**

egories, our method is more effective on the more challenging common and tail groups, suggesting that relational context is particularly beneficial when local visual evidence is ambiguous or insufficient.

Due to space limitations, additional qualitative comparisons and 3D open-vocabulary query results can be found in the supplementary material.

4.3 Zero-Shot Cross-Dataset Evaluation

To rigorously evaluate the cross-dataset generalization capabilities of our model, we conduct strict zero-shot inference on ScanNet++ [51] and Replica [39]. Specifically, our model is trained exclusively on the real-world ScanNet dataset and evaluated directly on the target datasets without any prior training or fine-tuning on their data.

Evaluation on ScanNet++. We first evaluate on ScanNet++, a benchmark of higher-quality reconstructed indoor scenes, under its standard 100-class setting. As shown in Tab. 4, *RelGraphOV* clearly outperforms representative open-vocabulary 3D baselines, indicating that the learned relational feature refinement generalizes beyond the ScanNet validation setting.

Evaluation on Replica. Beyond the real-world scans of ScanNet++, we further assess generalization to the synthetic, photorealistic Replica dataset, which exhibits a larger domain gap from the ScanNet training data. As shown in Tab. 5, RelGraphOV generalizes well to these unseen environments, achieving the highest overall mIoU (22.7%) and mAcc (32.7%) among the compared methods. The head/common/tail breakdown shows that our method improves rare-category mIoU, although OpenNeRF [10] obtains a higher tail mAcc. These results suggest that preserving CLIP-based semantic cues in the auxiliary stream helps cross-dataset generalization.

Table 5: Zero-Shot Evaluation on Replica [39]. The Zero-Shot column indicates whether a method is evaluated without any training or fine-tuning on Replica. † denotes results reproduced by us.

Method	Zero-Shot	*All*		*Head*		*Common*		*Tail*	
		mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
OpenScene-2D† [33]	✓	12.9	19.4	31.8	43.1	6.1	13.9	1.2	1.2
OpenScene-3D† [33]	✓	10.4	17.4	26.8	38.5	4.5	13.5	0.0	0.0
OpenNeRF [10]	✗	20.4	31.7	35.4	46.2	20.1	31.3	5.8	**17.6**
RelGraphOV (Ours)	✓	**22.7**	**32.7**	**38.4**	**55.2**	**22.4**	**33.7**	**7.3**	9.2

Table 6: Ablation Study. We validate the effectiveness of our relationship-aware 3D scene understanding framework and the dual-stream architecture.

Setting	ScanNetV2 [7]		ScanNet200 [37]	
	mIoU	mAcc	mIoU	mAcc
w/o scene graph	50.7	62.6	9.5	14.4
w/o relation guidance	55.5	70.7	11.0	20.1
w/o global feature	56.0	71.3	11.1	20.2
w/o dual graph fusion	57.9	73.0	12.7	22.9
Ours (full)	**58.4**	**73.4**	**14.5**	**26.1**

Table 7: Ablation Study. We validate the effectiveness of our hierarchical contrastive learning and dual-alignment strategy.

Setting	ScanNetV2 [7]		ScanNet200 [37]	
	mIoU	mAcc	mIoU	mAcc
w/o $\mathcal{L}_{\text{hie}}$	52.2	65.5	7.8	12.4
w/o $\mathcal{L}_{\text{reg}}$	54.9	71.7	10.9	20.1
w/o $\mathcal{L}_{\text{aux}}$	55.9	70.4	13.4	25.6
Ours (full)	**58.4**	**73.4**	**14.5**	**26.1**

4.4 Ablation Study & Analysis

Effectiveness of Dual-Stream Architecture and Relation-guided GAT. To validate the structural components of our framework, we conduct a multi-step ablation study (Tab. 6). First, to evaluate the role of message passing, we remove all edges, leaving only the initial node features (w/o scene graph). This degenerates our model into merely pooling LSeg features within isolated 3D object masks without any CLIP feature injection or inter-node communication, causing a clear performance drop and highlighting the need for contextual aggregation. Second, we verify the importance of edge-feature guidance by removing all edge features while preserving graph topology (w/o relation guidance). This reduction to a vanilla GAT leads to a noticeable drop, indicating that relationship semantics help filter geometric noise. Finally, to validate our dual-stream design, we evaluate a variant (w/o dual graph fusion) where LSeg and CLIP features are naively concatenated and processed in a single GNN stream. The resulting performance drop supports the "Feature Interference" hypothesis: mixing modalities in a single stream leads to cross-contamination. Our decoupled propagation and adaptive gated fusion are important for absorbing complementary semantics.

Global Information Injection. An ablation of the global information extraction and injection module is performed, with results in Tab. 6 showing its contribution. Removing it (w/o global feature) causes a clear performance drop, indicating its role in enhancing nodes' global contextual awareness. The module

integrates scene-level context to mitigate the limited receptive field of traditional GNNs and supports long-range dependency modeling in our architecture.

Hierarchical Contrastive and Dual-Alignment Learning Strategy. We evaluate our proposed loss scheme for aligning node features with textual embeddings (Tab. 7). First, removing the hierarchical contrastive loss ($\mathcal{L}_{hie}$) leads to a substantial performance drop, particularly on the challenging ScanNet200 benchmark. This indicates that $\mathcal{L}_{hie}$ is important for maintaining category-level distinctions and reducing inter-class confusion. Furthermore, ablating the generalized dual-alignment losses ($\mathcal{L}_{reg}$ and $\mathcal{L}_{aux}$) also leads to clear performance deterioration, highlighting the importance of preserving specific priors during feature refinement. Specifically, removing the main stream alignment (w/o $\mathcal{L}_{reg}$) degrades the LSeg prior and weakens segmentation boundaries. Similarly, removing the auxiliary stream alignment (w/o $\mathcal{L}_{aux}$) leads to semantic drift; without anchoring features to the initial CLIP space, the auxiliary stream loses open-vocabulary generalizability and fails to effectively supplement the main stream.

5 Conclusion

In this work, we presented *RelGraphOV*, a relationship-aware framework that incorporates environmental context into open-vocabulary 3D scene understanding via 3D scene graphs. We introduced a multi-view reasoning based scene graph construction pipeline to infer object relationships without manual annotations, and proposed an Adaptive Gated Dual-Stream Contextual GAT with hierarchical contrastive learning to enable robust contextual feature aggregation. Extensive experiments on real-world and synthetic benchmarks demonstrate state-of-the-art performance and strong cross-dataset generalization.

Limitations. Our current implementation uses fixed LSeg and CLIP feature backbones to isolate the effect of relational contextual refinement. Recent dense vision-language backbones may provide stronger initial open-vocabulary features, and integrating them into the same relational refinement framework is a promising direction for future work. In addition, our current formulation focuses on static reconstructed scenes. Extending *RelGraphOV* to dynamic scenes with evolving objects and relationships requires temporally consistent scene-graph construction and update mechanisms.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2024YFB4505500 & 2024YFB4505501. We also express our gratitude to all the anonymous reviewers for their professional and insightful comments.

References

1. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM. IEEE Trans. Robotics **37**(6), 1874–1890 (2021)

2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9630–9640. IEEE (2021)
3. Chang, H., Boyalakuntla, K., Lu, S., Cai, S., Jing, E.P., Keskar, S., Geng, S., Abbas, A., Zhou, L., Bekris, K.E., Boularias, A.: Context-aware entity grounding with open-vocabulary 3d scene graphs. In: Tan, J., Toussaint, M., Darvish, K. (eds.) Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA. Proceedings of Machine Learning Research, vol. 229, pp. 1950–1974. PMLR (2023)
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 1280–1289. IEEE (2022)
5. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 4113–4123. IEEE (2024)
6. Choy, C.B., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 3075–3084. Computer Vision Foundation / IEEE (2019)
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T.A., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2432–2443. IEEE Computer Society (2017)
8. Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., Nießner, M.: Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 4578–4587. Computer Vision Foundation / IEEE Computer Society (2018)
9. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: PLA: language-driven open-vocabulary 3d scene understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 7010–7019. IEEE (2023)
10. 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. In: International Conference on Learning Representations (2024)
11. Ghiasi, G., Gu, X., Cui, Y., Lin, T.: Scaling open-vocabulary image segmentation with image-level labels. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI. Lecture Notes in Computer Science, vol. 13696, pp. 540–557. Springer (2022)
12. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., Gan, C., de Melo, C.M., Tenenbaum, J.B., Torralba, A., Shkurti, F., Paull, L.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024. pp. 5021–5028. IEEE (2024)

13. Hu, J., Chen, X., Feng, B., Li, G., Yang, L., Bao, H., Zhang, G., Cui, Z.: CG-SLAM: efficient dense RGB-D SLAM in a consistent uncertainty-aware 3d gaussian field. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXV. Lecture Notes in Computer Science, vol. 15083, pp. 93–112. Springer (2024)
14. Hu, J., Mao, M., Bao, H., Zhang, G., Cui, Z.: CP-SLAM: collaborative neural point-based SLAM system. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
15. Huang, J., Zhang, H., Yi, L., Funkhouser, T.A., Nießner, M., Guibas, L.J.: Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 4440–4449. Computer Vision Foundation / IEEE (2019)
16. Jain, J., Li, J., Chiu, M., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 2989–2998. IEEE (2023)
17. Jiang, L., Shi, S., Schiele, B.: Open-vocabulary 3d semantic segmentation with foundation models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 21284–21294. IEEE (2024)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3992–4003. IEEE (2023)
19. Koch, S., Hermosilla, P., Vaskevicius, N., Colosi, M., Ropinski, T.: Lang3dsg: Language-based contrastive pre-training for 3d scene graph prediction. In: International Conference on 3D Vision, 3DV 2024, Davos, Switzerland, March 18-21, 2024. pp. 1037–1047. IEEE (2024)
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/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 14183–14193. IEEE (2024)
21. Koch, S., Wald, J., Colosi, M., Vaskevicius, N., Hermosilla, P., Tombari, F., Ropinski, T.: Relationfield: Relate anything in radiance fields. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 21706–21716. Computer Vision Foundation / IEEE (2025)
22. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: Mseg: A composite dataset for multi-domain semantic segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 2876–2885. Computer Vision Foundation / IEEE (2020)
23. Lee, J., Park, C., Choe, J., Wang, Y.F., Kautz, J., Cho, M., Choy, C.B.: Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 14089–14101. Computer Vision Foundation / IEEE (2025)

24. Li, B., Weinberger, K.Q., Belongie, S.J., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
25. Li, J., Saltori, C., Poiesi, F., Sebe, N.: Cross-modal and uncertainty-aware agglomeration for open-vocabulary 3d scene understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 19390–19400. Computer Vision Foundation / IEEE (2025)
26. Li, R., Zhang, Z., He, C., Ma, Z., Patel, V.M., Zhang, L.: Dense multimodal alignment for open-vocabulary 3d scene understanding. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIX. Lecture Notes in Computer Science, vol. 15107, pp. 416–434. Springer (2024)
27. Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M., Darrell, T., Yala, A., Cui, Y.: Describe anything: Detailed localized image and video captioning. CoRR **abs/2504.16072** (2025)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII. Lecture Notes in Computer Science, vol. 15105, pp. 38–55. Springer (2024)
29. Lu, J., Deng, J.: Relation3d : Enhancing relation modeling for point cloud instance segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 8889–8899. Computer Vision Foundation / IEEE (2025)
30. Naeem, M.F., Xian, Y., Zhai, X., Hoyer, L., Gool, L.V., Tombari, F.: SILC: improving vision language pretraining with self-distillation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXI. Lecture Notes in Computer Science, vol. 15079, pp. 38–55. Springer (2024)
31. Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F.: Mix3d: Out-of-context data augmentation for 3d scenes. In: International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021. pp. 116–125. IEEE (2021)
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024)
33. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.A.: Openscene: 3d scene understanding with open vocabularies. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 815–824. IEEE (2023)
34. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang,

P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025)

35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
36. 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)
37. Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII. Lecture Notes in Computer Science, vol. 13693, pp. 125–141. Springer (2022)
38. Schult, J., Engelmann, F., Kontogianni, T., Leibe, B.: Dualconvmesh-net: Joint geodesic and euclidean convolutions on 3d meshes. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 8609–8619. Computer Vision Foundation / IEEE (2020)
39. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
40. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
41. Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.: Tangent convolutions for dense prediction in 3d. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 3887–3896. Computer Vision Foundation / IEEE Computer Society (2018)
42. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 3960–3969. Computer Vision Foundation / IEEE (2020)
43. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
44. Wang, Y., Jia, B., Zhu, Z., Huang, S.: Masked point-entity contrast for open-vocabulary 3d scene understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 14125–14136. Computer Vision Foundation / IEEE (2025)

45. 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: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 21560–21569. IEEE (2023)
46. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
47. Wu, S., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegraphfusion: Incremental 3d scene graph prediction from RGB-D sequences. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 7515–7525. Computer Vision Foundation / IEEE (2021)
48. Yan, M., Zhang, J., Zhu, Y., Wang, H.: Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 28274–28284. IEEE (2024)
49. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. CoRR **abs/2310.11441** (2023)
50. Yang, J., Ding, R., Deng, W., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 19823–19832. IEEE (2024)
51. Yeshwanth, C., Liu, Y., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 12–22. IEEE (2023)
52. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 28202–28211. IEEE (2024)
53. Zhang, X., Bao, C., Chen, Y., Zhai, H., Dong, Y., Bao, H., Cui, Z., Zhang, G.: Atlasgs: Atlanta-world guided surface reconstruction with implicit structured gaussians. Advances in Neural Information Processing Systems **38**, 15556–15580 (2026)
54. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)