

LEGO: Leveled Language Gaussian Splatting

Yuning Peng¹, Haiping Wang^{1,2,†}, Yuan Liu², Yipeng Lu¹, Zhen Dong¹,
and Bisheng Yang¹

¹ Wuhan University

{yuningpeng,hpwang,luyipeng,dongzhenwu,bshyang}@whu.edu.cn

² Hong Kong University of Science and Technology

yuanly@connect.hku.hk

Abstract. We introduce LEGO for advanced open-vocabulary scene understanding. Beyond basic concept recognition, its core innovation lies in capturing the intrinsic semantic hierarchies within the scene, such as the "flowerpot $\rightarrow$ bouquet $\rightarrow$ bud $\rightarrow$ petal" lineage. While foundation models like SAM can identify multi-granular structures in 2D, their partitions are strictly perspective-bound and lack cross-view consensus. LEGO self-adaptively re-grades volatile multi-view SAM granularities into a unified, 3D-consistent hierarchy. This provides precise supervision for the structurally coherent, multi-level segmentation of 3D scenes. By grounding these segments with CLIP embeddings, LEGO recovers open-vocabulary semantic logic across hierarchical levels. Furthermore, by incorporating spatial relationships, we elevate these segments into level-wise language scene graphs, effectively empowering Large Language Models to perform complex, context-aware spatial reasoning and precise visual grounding. Experimental results demonstrate that LEGO establishes new state-of-the-art performance across both promptable and open-vocabulary 3D segmentation benchmarks, exhibiting advanced hierarchical scene decomposition and context-aware spatial reasoning. Project page: <https://pz0826.github.io/LEGO-Webpage/>

Keywords: 3D Semantic Hierarchy · Open-vocabulary Scene Understanding · 3D Scene Graph

1 Introduction

As 3D scene segmentation evolves from closed-set taxonomies to open-vocabulary understanding, a critical aspect is that *real-world semantics are inherently hierarchical*. When humans perceive a scene, objects can be interpreted at various levels based on their semantic context, such as "flowerpot $\rightarrow$ bouquet $\rightarrow$ bud $\rightarrow$ petal". Therefore, a robust 3D open-vocabulary system must not only recognize arbitrary concepts but also explicitly construct the hierarchical semantic structures underlying the 3D space.

[†] Corresponding author.



Fig. 1: LEveled Gaussian splatting for Open-vocabulary scene understanding (LEGO). By parsing scenes into a unified, level-wise semantic hierarchy, LEGO establishes a robust foundation for multi-granular open-vocabulary understanding, seamlessly enabling automatic decomposition and context-aware spatial reasoning.

Segment Anything Model (SAM) [22] enables hierarchical 2D segmentation through three granularities: whole, part, and subpart. While lifting these masks as 3D levels is a seemingly direct path to building 3D hierarchies, existing methods face two fundamental limitations.

View-Dependency in Granularity-based Distillation. The first category of work [6, 8, 9, 37, 53] directly maps SAM’s 2D granularities to 3D hierarchies. However, SAM labels are inherently view-dependent, determined by an object’s relative granularity in the 2D image plane rather than its intrinsic 3D structure. For example, as shown in Fig. 2(a), a flower may be segmented into individual petals in a close-up but a complete bud from a distance at the same granularity. Distilling such inconsistent labels into 3D space leads to granularity blurring, where the model fails to assign a stable hierarchical rank to the same physical entity.

Semantic-Scale Gap in Scale-based Distillation. To ensure view invariance, a second category of work [4, 20, 21] groups 3D segments based on absolute physical scales. However, relying on a manual scale parameter to generate segments leads to semantic-scale decoupling, as semantic entities often exhibit high intra-class scale variance. For instance, as shown in Fig. 2(b), flowers within a bouquet vary greatly in size. Under a chosen scale parameter, a large flower might be over-segmented into petals, while a smaller flower at the

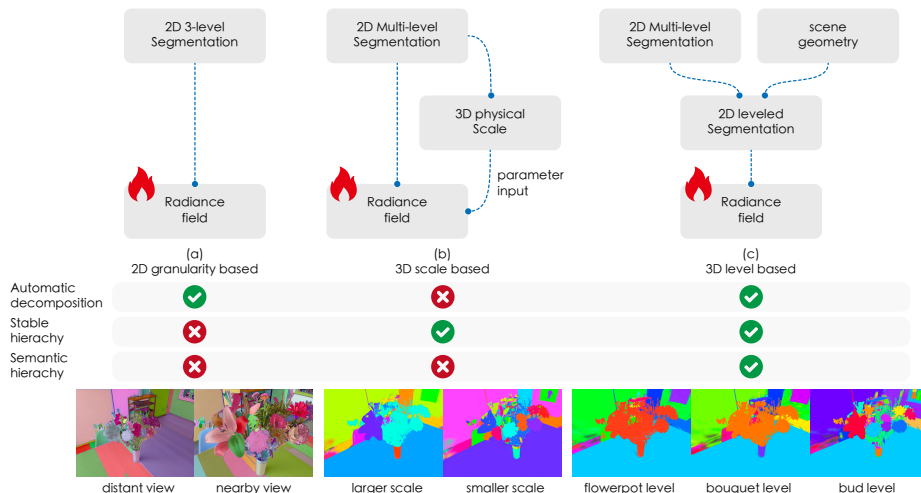


Fig. 2: Comparison of 3D hierarchical segmentation paradigms. (a) **2D Granularity-based** methods lift view-dependent masks, leading to inconsistent 3D hierarchies across different camera distances (*e.g.*, petals in a nearby view vs. a whole bud in a distant view). (b) **3D Scale-based** methods rely on absolute physical sizes, suffering from a semantic-scale gap. Due to intra-class scale variance, a single scale parameter may over-segment a large flower into petals while keeping smaller ones intact. (c) **Our 3D Level-based paradigm** automatically discovers a stable and consistent semantic hierarchy, grouping entities by their true structural compositional rank (*e.g.*, flowerpot $\rightarrow$ bouquet $\rightarrow$ bud) regardless of viewing distance or absolute physical size.

same semantic level remains intact. Consequently, these methods require exhaustive per-instance tuning rather than discovering a consistent semantic hierarchy, hindering the automated reasoning essential for open-world understanding.

Our Insight: From Granularities and Scales to Structural Levels.

To build an accurate 3D semantic hierarchy, we argue that the system must shift from view-dependent *granularities* and manual physical *scales* to intrinsic **structural levels**—a consistent compositional rank that is strictly invariant to both viewing distances and absolute physical sizes (*e.g.*, all flowers in a bouquet belong to the "bud" level, regardless of how large they are or how close the camera is).

To achieve this, we present LEGO (**LE**veled **G**aussian splatting for **O**pen-vocabulary understanding). Our fundamental insight is that while SAM’s 2D granularity labels are volatile across viewpoints, they collectively serve as partial glimpses into a scene’s underlying semantic hierarchy. For example, a distal viewpoint might perceive "flowerpot $\rightarrow$ bouquet $\rightarrow$ bud" as its three granularities, while a proximal viewpoint captures "bouquet $\rightarrow$ bud $\rightarrow$ petal". Although their relative definitions shift, these observations jointly encapsulate the complete semantic lineage: "flowerpot $\rightarrow$ bouquet $\rightarrow$ bud $\rightarrow$ petal". The core of LEGO lies in its ability to adaptively re-grade all multi-view masks covering

the same 3D region into a unified, level-wise structure, effectively "stitching" fragmented 2D granularities into a coherent 3D hierarchy.

Specifically, we aggregate multi-view SAM masks and estimate their corresponding 3D physical scales. Since masks representing the same semantic entity exhibit consistent spatial dimensions across viewpoints, we utilize these physical scales as an *initial proxy* to establish the semantic hierarchy. By clustering masks based on spatial co-visibility and 3D scale, we effectively discretize the continuous scales into distinct structural levels, thereby manifesting the implicit 3D hierarchy of the scene. As illustrated in Fig. 2(c), our 3D *level-based* paradigm provides accurate, natively consistent supervision for hierarchical Gaussian scene segmentation, effortlessly parsing the scene into stable tiers.

To this end, we ground these segments with optimal-view CLIP features, establishing a *multi-level, open-vocabulary hierarchy* for scene objects. We also show that, by further integrating spatial context of objects, we elevate these segments into *level-wise language scene graphs*, effectively empowering Large Language Models (LLMs) to perform complex, context-aware spatial reasoning.

2 Related Work

2.1 Segmentation in Radiance Fields

The emergence of Neural Radiance Fields (NeRF) [31] and 3D Gaussian Splatting (3DGS) [19] has revolutionized 3D scene representation. While many methods [28, 50, 51, 58, 59] successfully lift 2D instance masks to 3D Radiance Field, they predominantly focus on object-centric tasks, overlooking fine-grained part-level details essential for comprehensive scene understanding.

To capture these finer details, some works have ventured into part-level segmentation. However, addressing the inherent ambiguity of part decomposition remains a challenge. One line of work [4, 21, 52] introduces global regulatory parameters—such as physical scale or feature similarity thresholds—to resolve this ambiguity. Although capable of separating parts, these methods often require manual tuning of these global parameters to match the specific granularity of a target object, limiting their ability to handle complex scenes with diverse intra-class scales. Another stream of research [16, 54, 55] adopts a spatial interactive paradigm, relying on user clicks or 2D bounding boxes as initialization cues. While effective for human-in-the-loop applications, these methods lack the autonomy required for scalable, automatic processing. In contrast, LEGO achieves native, autonomous hierarchical segmentation by discovering consistent structural levels without predefined scale/similarity priors or manual guidance.

2.2 Open-Vocabulary 3D Understanding

To equip 3D representations with semantic understanding, early methods distilled patch-level CLIP embeddings [20, 27, 42, 60] or leveraged dense features from other 2D vision encoders [3, 11, 23, 33, 35, 38, 41], but often struggled with blurred feature boundaries or long-tail categories.

With the advent of the Segment Anything Model (SAM), many works [17, 18, 24, 34, 49, 51, 56, 57] utilize SAM masks to aggregate CLIP features, significantly improving boundary adherence. However, these methods primarily operate at the object level. While some attempts [6, 8, 9, 36, 37] have explored part-level lifting, they often encounter severe context deficiency. This is particularly problematic for low-texture components, which lack distinctive visual features when isolated from the whole object, leading to incorrect feature extraction. To mitigate this, several approaches [2, 15, 43, 44] tried to incorporate context by fusing features from surrounding or parent objects. Nevertheless, this strategy inevitably leads to feature averaging and blurring, resulting in semantic ambiguity where distinct parts become indistinguishable from their neighbors. Diverging from these, LEGO employs a structurally decoupled feature space. By optimizing hierarchical levels independently, we strictly prevent cross-level semantic entanglement, extracting pristine features for robust, context-aware part-level understanding.

2.3 Structured Scene Modeling

Beyond dense semantic fields, structuring 3D scenes into graphs provides a more interpretable representation. Object-level scene graphs [13, 26, 46, 47] abstract scenes into nodes representing objects and edges representing spatial relationships. While effective for navigation, they lack the granularity to describe intra-object structures. To incorporate hierarchy, HOV-SG [48] constructs a multi-layer floor-room-object graph but stops short of decomposing objects into parts. Recently, a few works [9, 53] have attempted part-level hierarchical modeling. However, these methods are fundamentally constrained by SAM’s architecture, which only provides three predefined output levels (whole, part, subpart). This rigid limitation prevents them from adapting to the varied complexity of real-world scenes. LEGO transcends this limitation by proposing an adaptive decomposition strategy. Unconfined by view-dependent 2D granularities, LEGO constructs a flexible, multi-layer 3D scene graph, effectively empowering Large Language Models (LLMs) to perform complex, context-aware spatial reasoning.

3 Method

Given a set of multi-view RGB images $I = \{I_1, I_2, \dots, I_V\}$ of a scene, LEGO aims to output a 3D Gaussian field enriched with hierarchical semantic features, allowing a multi-level scene graph construction for LLM-based scene reasoning.

Specifically, LEGO first initializes a 3D Gaussian field [45] $\mathcal{P}$ using the geometric prior reconstructed by MAST3R-SfM [10]. Then, in Sec. 3.1, we re-grade multi-view SAM masks into 3D-consistent leveled masks. These masks then facilitate the hierarchical segmentation of $\mathcal{P}$ in Sec. 3.2. Finally, in Sec. 3.3, we embed CLIP language features into the resulting segments and organize them into a scene graph based on their spatial relationships, empowering Large Language Models to perform complex spatial grounding tasks.

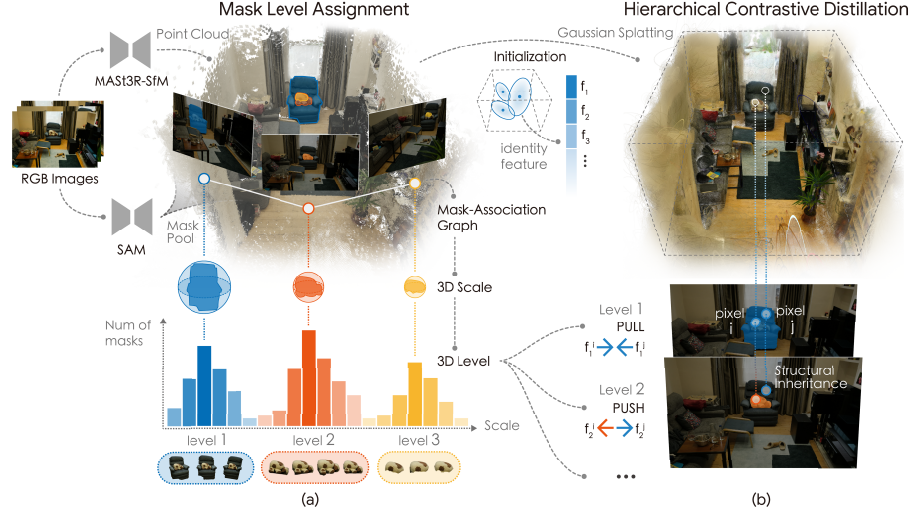


Fig. 3: LEGO pipeline. (a) LEGO first lifts multiview SAM masks into a 3D coherent scale space to assign them view-consistent structural levels (Sec. 3.1). (b) Guided by these level-masks, a hierarchical contrastive distillation process optimizes the decoupled identity features of the Gaussian field (Sec. 3.2).

3.1 3D Semantic Hierarchy from SAM Masks

In this section, we re-grade multi-view, multi-granularity SAM masks into a consistent 3D semantic hierarchy. Specifically, each mask is assigned a structural level label that characterizes its relative semantic scale within its local region.

Mask Lifting and 3D Scale Estimation. We first aggregate multi-view SAM masks into a global pool $\mathcal{M}$. Using camera poses and scene representation $\mathcal{P}$ from MAST3R-SfM [10], we establish a pixel-to-point mapping $\mathcal{F} : u \mapsto \mathbf{p}$. For each 2D mask $m_i \in \mathcal{M}$, we lift its constituent pixels to 3D points $\mathcal{P}_{m_i} = \{\mathcal{F}(u) \mid u \in m_i\}$. The physical scale s_i of mask m_i is derived as the effective spatial diameter based on the point set’s standard deviation: $s_i = 2\sqrt{\sum_{d \in \{x,y,z\}} \text{std}(\mathcal{P}_{m_i,d})^2}$.

Local Peak-based Mask Level Assignment. We determine the hierarchical level of m_i by situating it within its local structural context. Specifically, for each target mask m_i , we first identify its spatially neighboring masks $\mathcal{N}(m_i)$ that share co-visible 3D regions with it.

Then, we perform peak detection on the histogram of $\mathcal{N}(m_i) \cup \{m_i\}$ scales to self-adaptively identify L prominent peaks $\mathcal{K}_i = \{p_1, p_2, \dots, p_L\}$, sorted by scale from coarse to fine. These peaks represent the underlying semantic hierarchy (e.g., p_1 for the root object, p_L for the finest sub-parts) present in this local 3D region. Finally, the target mask m_i is assigned a discrete level l_i based on which peak its own scale s_i aligns with:

$$l_i = \arg \min_{l \in \{1, 2, \dots, L\}} |s_i - p_l|. \quad (1)$$

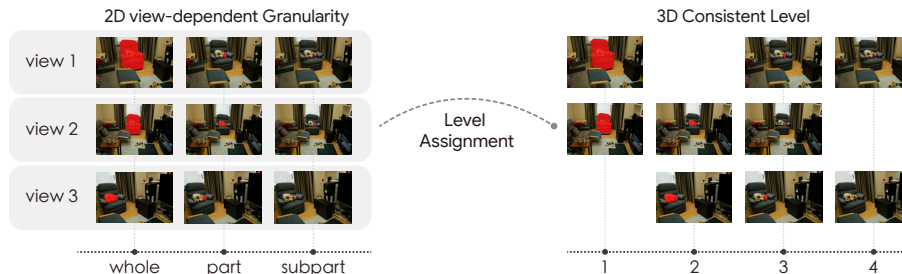


Fig. 4: Illustration of mask level assignment. Our mechanism discovers the complete 3D hierarchical structure implicit in multi-view 2D masks, assigning each mask to its corresponding view-consistent level.

Through this per-mask assignment, as illustrated in Fig. 4, LEGO ensures that each mask is "self-graded" into a view-consistent local 3D semantic hierarchy, effectively resolving the ambiguity of 2D granularity by leveraging the collective consensus of its co-visible neighbors.

Level-wise Mask Dense Indicator. Once each mask is assigned to a structural level, a naive supervision strategy would be to optimize level- k segmentation using only the masks strictly graded at that level. However, because our reconstructed 3D hierarchy contains significantly more levels than SAM’s three 2D granularities, the re-graded SAM masks for any single level are inherently sparse. Supervision solely with these sparse masks leads to optimization instability and structural artifacts in level-wise 3D segmentation. To resolve this, we design several strategies to densify the level-wise supervision by propagating mask labels through the hierarchy.

In implementation, as illustrated in Fig. 3(b), rather than explicitly densifying levelwise mask maps, we derive an equivalent but highly parallelizable pixel-wise indicator $\mathbb{I}$: for any pixel pair (i, j) in view v , $\mathbb{I}_k(i, j)$ determines whether they belong to the same entity mask at level k , formulated as

$$\mathbb{I}_k(i, j) = \begin{cases} 1 & \text{if } \mathcal{M}_k(i) \cap \mathcal{M}_k(j) \neq \emptyset \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where $\mathcal{M}_k(i)$ denotes the set of active masks covering pixel i at level k .

Notably, the construction of $\mathbb{I}$ is non-trivial, requiring strict adherence to the following hierarchical axioms to maintain structural integrity:

- **Monotonicity:** Supervisory signals are strictly non-increasing as the hierarchy deepens; pixel pairs may split into sub-parts as granularity increases, but they can never merge at finer levels.
- **Recursive Inclusion:** Identity at a fine-grained level strictly necessitates identity at all coarser parent levels, ensuring nested structural consistency across the field.
- **Structural Inheritance:** To mitigate mask sparsity and prevent optimization collapse, pixel lacking a valid mask at the current target level inherit

labels from their nearest valid coarser level, ensuring dense and stable supervision at every layer.

These principles yield a conflict-free and hierarchically-consistent dense supervision by pixel-pair parallelism, stabilizing efficient level-wise segmentation in the following section. More details are provided in the Supplementary Material.

3.2 Segmentation with Level-wise Semantic Hierarchy

In this section, we describe the supervision of the hierarchical segmentation for Gaussian field $\mathcal{P}$ using the level-wise topological constraints established above. Inspired by SAGA [4], we equip each Gaussian primitive with a learnable identity embedding. By enforcing the rendered feature distributions to align with our level-wise indicator constraints, the model learns to capture the spatial and semantic distinctness of different regions. To support this hierarchical supervision, we embed a decoupled feature space where each structural level is optimized independently without cross-level interference.

Level Feature Embedding. To represent a multi-granular hierarchy without inter-level interference, we assign each Gaussian primitive in $\mathcal{P}$ a composite identity feature $\mathbf{f} \in \mathbb{R}^{L \times d}$. We structurally decouple the feature space into L independent d -dimensional subspaces, where the feature for level k is directly routed to its k -th row vector $\mathbf{f}_k \in \mathbb{R}^d$. This disjoint design prevents semantic entanglement, allowing the model to optimize the feature representation of each hierarchical level independently.

Level-wise Feature Learning. We formulate a contrastive distillation loss to dynamically align the Gaussian features with the dense hierarchical supervision. For a given training viewpoint v and target level k , we first render the level-specific primitive features $\mathbf{f}_k$ to a 2D feature map $\mathbf{F}_{v,k}$.

For a sampled pixel pair (i, j) at level k , the contrastive objective is defined as:

$$\mathcal{L}_{contrast} = w_k^{i,j} (1 - 2 \cdot \mathbb{I}_k(i, j)) \cos(\mathbf{F}_{v,k}^i, \mathbf{F}_{v,k}^j) \quad (3)$$

where $\mathbf{F}_{v,k}^i$ is the rendering feature of pixel i . The indicator function $\mathbb{I}_k(i, j)$, as formulated in Sec. 3.1, dynamically dictates the pairing relationship based on the intersection of the active mask sets $\mathcal{M}_k(i)$ and $\mathcal{M}_k(j)$.

To prevent large-scale entities from dominating the gradient, we employ a scale-balancing weight $w_k^{i,j} \propto (\mathcal{A}_k^i \cdot \mathcal{A}_k^j)^{-1}$, where $\mathcal{A}_k^i$ denotes the area of the active mask governing pixel i at level k . This ensures that small objects contribute equally to the optimization as massive ones within each level. Furthermore, as cosine similarity only constrains feature orientation, we impose an auxiliary L_2 penalty on positive pairs to enforce strict feature identity:

$$\mathcal{L}_{pos} = \mathbb{I}_k(i, j) \|\mathbf{F}_{v,k}^i - \mathbf{F}_{v,k}^j\|_2^2 \quad (4)$$

Moreover, since rendered pixels are weighted accumulations of multiple Gaussians, the model might exploit view-dependent shortcuts that fail to represent

consistent 3D geometry. We counteract this by constraining both 3D primitive features $\mathbf{f}_k$ and rendered 2D features $\mathbf{F}_{v,k}^i$ to a unit hypersphere:

$$\mathcal{L}_{3d} = |1 - \|\mathbf{f}_k\|_2|, \quad \mathcal{L}_{2d} = |1 - \|\mathbf{F}_{v,k}^i\|_2| \quad (5)$$

This ensures that all Gaussians along a single ray possess identical, consistent features, preventing feature decay during blending for a truly view-invariant 3D semantic field.

Hierarchical Scene Segmentation. After convergence of the above feature learning, LEGO performs hierarchical segmentation through a top-down recursive clustering strategy, naturally representing the scene as a tree-structured semantic hierarchy. Starting from the root node (the entire scene), we iteratively partition parent nodes into child clusters. Specifically, for a parent node at level $k - 1$, we apply HDBSCAN [29] to its constituent Gaussian primitives based on their level- k features $\mathbf{f}_k$. This process continues recursively until the maximum hierarchy depth L is reached. By leveraging the decoupled nature of our feature subspaces, this approach ensures that the resulting segments are not merely isolated clusters; instead, they are structurally nested to faithfully reflect the multi-granular semantic hierarchy distilled from the leveled masks.

3.3 Level-wise Language Scene Graph

In this section, we endow the hierarchical segments with CLIP embeddings for open-vocabulary understanding. To further facilitate complex spatial reasoning, we incorporate spatial context to synthesize these language-grounded segments into level-wise scene graphs. This structured representation captures both hierarchical containment (part-whole) and spatial proximity, effectively empowering the model to resolve complex, context-aware queries such as "Find the handle of the pitcher beside the rolling pin."

View-Aware Language Grounding. When embedding image CLIP features into a 3D segment, to avoid semantic noise introduced by occlusions in a naive multi-view average, we propose an Optimal View Selection (OVS) strategy. For segment c , the observation quality score $\mathcal{S}(c, v)$ for view I_v is:

$$\mathcal{S}(c, v) = \underbrace{\left(\frac{|c_{\text{vis}}^v|}{|c|} \right)}_{\text{3D Visibility}} \cdot \underbrace{\left(\frac{|m_c^v|}{|H \times W|} \right)}_{\text{2D Coverage}} \cdot \underbrace{\text{IoU}(m_c^v, m^*)}_{\text{Semantic Alignment}} \quad (6)$$

where $|c|$ denotes the total number of Gaussians in c , $|c_{\text{vis}}^v|$ is the subset of those Gaussians successfully mapped to valid pixels in frame I_v , m_c^v is the projected 2D mask of c , m^* represents the best-aligned SAM mask in view v , and $H \times W$ is the total pixel area of the image. "3D Visibility" penalizes segments that are heavily occluded in the current frame, "2D Coverage" prioritizes viewpoints where the projected segment occupies a significant portion of the view, "Semantic Alignment" ensures the projected segment mask m_c^v aligns well with the 2D SAM prior m^* .

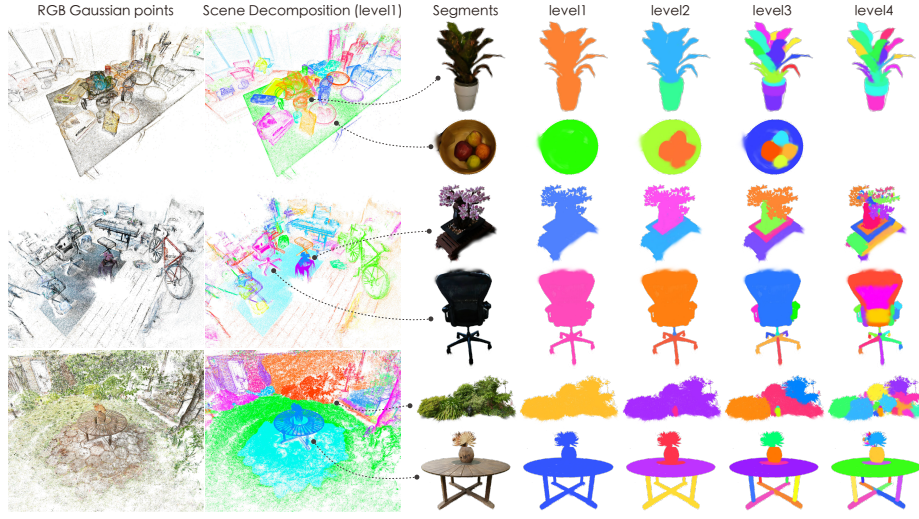


Fig. 5: Visualization of automatic hierarchical scene decomposition. Benefiting from our level-wise design, LEGO captures arbitrarily deep hierarchical structures. Here, selected entities are progressively parsed into extreme fine-grained sub-parts, while maintaining strict structural nestedness across multiple levels.

Then, we retrieve the top- τ frames that maximize $\mathcal{S}(c, v)$. The corresponding high-quality SAM masks m^* are utilized to crop the input images and extract CLIP embeddings. The final open-vocabulary feature $\mathbf{E}_c$ is obtained via average pooling of these top- τ embeddings. This view-aware strategy ensures pristine, high-fidelity semantic representations while drastically reducing computational overhead.

From Segments to Scene Graph. To move beyond a simple semantic tree and capture the full spatial context, we define two types of edges:

- Hierarchical Edges ($\mathcal{E}_{hier}$): These encode the vertical "part-whole" relationships established during the hierarchical scene segmentation process.
- Spatial Adjacency Edges ($\mathcal{E}_{adj}$): These represent the horizontal context. We establish an edge between two independent nodes if their 3D bounding spheres intersect. This provides the topological foundation for reasoning about spatial relations like "on" or "beside".

The resulting scene graph enables sophisticated tasks like Chain-of-Retrieval (CoR). When faced with a complex query (*e.g.*, "Find the handle of the pitcher beside the rolling pin"), an LLM agent parses it into a coarse-to-fine sequence (rolling pin $\rightarrow$ pitcher $\rightarrow$ handle). LEGO then executes a search over the graph topology, using coarser objects as spatial anchors to retrieve fine-grained parts through semantic matching and spatial relationship verification.

Table 1: Results on NVOS dataset.

Method	mIoU(%)	mAcc(%)
NVOS [39]	70.1	92.0
ISRF [12]	83.8	96.4
SA3D [7]	90.3	98.2
OmniSeg3D [52]	91.7	98.4
GS-Grouping [51]	85.6	97.3
FlashSplat [40]	91.8	<u>98.6</u>
iSegMan [55]	92.0	98.4
COB-GS [54]	92.1	<u>98.6</u>
SA3D-GS [5]	92.2	98.5
SAGA [4]	<u>92.6</u>	<u>98.6</u>
LEGO(Ours)	94.2	98.7

Table 2: Results on SPIn-NeRF dataset.

Method	mIoU(%)	mAcc(%)
MVSeg [32]	90.9	98.9
SA3D [7]	92.4	98.9
OmniSeg3D [52]	94.3	99.3
GS-Grouping [51]	86.5	98.9
SAGD [14]	90.0	98.7
iSegMan [55]	92.4	99.1
SA3D-GS [5]	93.2	99.1
SAGA [4]	93.4	<u>99.2</u>
LEGO(Ours)	<u>94.2</u>	99.3

4 Experiments

Tasks and Datasets. We comprehensively evaluate the performance of LEGO on both segmentation and understanding tasks:

Promptable Segmentation. We conduct experiments on the NVOS [39] and SPIn-NeRF [32] datasets. The NVOS dataset, derived from the established LLFF [30] dataset, comprises 7 diverse real-world scenes with complex geometric structures. The SPIn-NeRF dataset integrates subsets from several existing NeRF benchmarks, providing high-quality real-world captures designed to evaluate precise multi-view segmentation.

Open-vocabulary Understanding. We utilize two real-world datasets: the LERF-OVS dataset [20] (extended with LangSplat [37] annotations) and Mip-NeRF 360 dataset [1] (annotated by GAGS [36]). The LERF dataset contains over 3,700 high-resolution images captured by mobile phones across 14 diverse scenes. LangSplat provides text descriptions and corresponding multi-view segmentation masks within four scenes (“*ramen*”, “*waldo_kitchen*”, “*teatime*”, and “*figures*”) to assess text-based localization and segmentation. The Mip-NeRF 360 dataset offers unbounded and highly intricate indoor and outdoor scenes. Following GAGS, we evaluate text-based localization and segmentation in four scenes (“*room*”, “*counter*”, “*garden*”, and “*bonsai*”).

For open-vocabulary evaluations with single-word or phrase queries, we bypass the LLM scene graph (Sec. 3.3) and simply perform standard CLIP feature similarity across the first three hierarchical levels to retrieve the target object.

Evaluation Metrics. We evaluate promptable segmentation using standard mean Intersection over Union (mIoU) and mean Accuracy (mAcc). For open-vocabulary tasks, we assess semantic grounding via localization mAcc (predictions within ground-truth bounding boxes) and text-queried segmentation mIoU.

4.1 Results

Promptable Segmentation. Tables 1 and 2 report quantitative results on NVOS and SPIn-NeRF. LEGO achieves superior or highly competitive perfor-

Table 3: Open-vocabulary 3D understanding results on LERF-OVS dataset.

Method	3D Location mAcc(%)					3D Segmentation mIoU(%)				
	<i>Ramen</i>	<i>Teatime</i>	<i>Kitchen</i>	<i>Figurines</i>	Overall	<i>Ramen</i>	<i>Teatime</i>	<i>Kitchen</i>	<i>Figurines</i>	Overall
LEGaussian [42]	69.0	79.7	63.6	57.1	67.4	20.2	32.3	22.3	23.4	24.6
GS-Grouping [51]	32.4	69.5	50.0	44.6	49.1	26.4	54.0	31.3	34.6	36.6
GOI [38]	56.3	67.8	68.2	44.6	59.2	33.7	55.8	54.5	23.9	42.0
GAGS [36]	69.0	88.1	<u>90.9</u>	78.6	81.7	46.8	60.3	55.8	53.6	54.1
LangSplat [37]	73.2	88.1	95.5	<u>80.4</u>	<u>84.3</u>	51.2	65.1	44.5	44.7	51.4
Occam’s LGS [8]	74.7	<u>93.2</u>	81.8	<u>80.4</u>	82.5	51.0	70.2	<u>65.3</u>	58.6	61.3
LangSplatV2 [25]	74.7	<u>93.2</u>	86.4	82.1	84.1	51.8	<u>72.2</u>	59.1	56.4	59.9
LaGa [6]	<u>76.1</u>	84.8	81.8	78.6	80.3	<u>55.6</u>	70.9	65.6	64.1	<u>64.0</u>
LEGO(Ours)	87.3	94.9	<u>90.9</u>	<u>80.4</u>	88.4	67.5	81.0	62.7	<u>62.3</u>	68.4

Table 4: Open-vocabulary 3D understanding results on Mip-NeRF 360 dataset.

Method	3D Location mAcc(%)					3D Segmentation mIoU(%)				
	<i>Room</i>	<i>Counter</i>	<i>Garden</i>	<i>Bonsai</i>	Overall	<i>Room</i>	<i>Counter</i>	<i>Garden</i>	<i>Bonsai</i>	Overall
LEGaussian [42]	55.2	73.0	71.4	61.1	65.2	25.5	35.3	33.2	22.3	29.1
GS-Grouping [51]	<u>79.3</u>	56.8	57.1	66.7	65.0	54.4	47.7	40.4	54.1	49.1
GOI [38]	69.0	67.6	66.7	72.2	68.9	60.3	46.6	59.8	67.3	58.5
GAGS [36]	93.1	97.3	<u>81.0</u>	<u>83.3</u>	<u>88.7</u>	65.2	61.1	61.2	70.5	64.5
LangSplat [37]	75.9	91.9	52.4	72.2	73.1	51.2	64.7	49.8	53.0	54.7
Occam’s LGS [8]	93.1	97.3	57.1	72.2	79.9	<u>69.2</u>	<u>72.9</u>	50.9	62.2	63.8
LangSplatV2 [25]	75.9	91.9	85.7	<u>83.3</u>	84.2	64.3	75.1	<u>65.0</u>	<u>73.1</u>	<u>69.4</u>
LaGa [6]	93.1	<u>94.6</u>	71.4	<u>83.3</u>	85.6	72.6	72.1	42.1	68.0	63.7
LEGO(Ours)	93.1	97.3	85.7	94.4	92.6	67.3	71.3	71.0	82.4	73.0

mance compared to prior methods, notably securing non-trivial margins (up to +1.6% mIoU on NVOS) on these heavily saturated benchmarks. By explicitly grouping masks into distinct hierarchical levels and optimizing disjoint sub-features, LEGO insulates the representation from the noise of low-quality priors. This eliminates cross-level feature entanglement, resulting in a purer and strictly consistent feature field at every level.

Open-Vocabulary Understanding. As shown in Tables 3 and 4, LEGO consistently achieves state-of-the-art performance, yielding average improvements of +4.1% localization mAcc and +4.4% segmentation mIoU on LERF-OVS, alongside +3.9% mAcc and +3.6% mIoU on Mip-NeRF 360, achieving top or near-top performance across the vast majority of scenes.

Crucially, LEGO excels in segmenting extreme fine-grained sub-parts. In Fig. 6, while baselines only locate prominent entities like “eggs” in the *Ramen* scene, LEGO cleanly isolates fine-grained “corn” and “onion segments”, yielding a +11.2% mAcc and +11.9% mIoU boost over the best baseline. Similarly, in the *Bonsai* scene, LEGO successfully segments tiny objects (*e.g.*, “pedal of piano”) and preserves exquisite boundaries for topologically complex objects with intricate edges (*e.g.*, “LEGO bonsai”), whereas alternative methods fail or erroneously highlight the entire parent object due to feature blending.

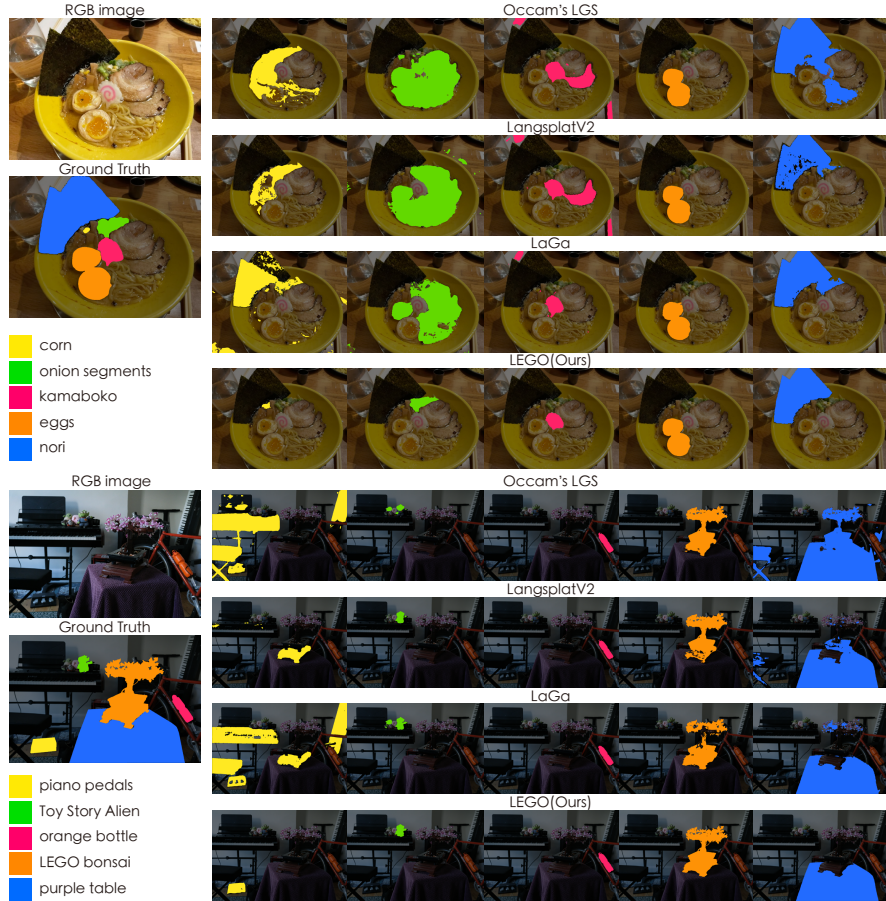


Fig. 6: Open-vocabulary segmentation results. While other methods struggle with severe feature blurring, LEGO cleanly isolates extremely fine-grained entities like “corn” and “onion segments”, and preserves exquisite boundaries for “LEGO bonsai”.

LLM-Guided Complex Scene Reasoning. We further evaluate LEGO’s capability to parse compositional prompts that require deep hierarchical and relational reasoning—a known failure point for standard CLIP-based fields due to the “bag-of-words” effect. As illustrated in Fig. 7, we apply the LLM-guided Chain-of-Retrieval (CoR) over our constructed scene graph. For highly intricate spatial queries (e.g., “Find the mug on a cardboard box next to the bicycle”) or extreme fine-grained part localization (e.g., “Segment the logo on Pringles chip”), LEGO seamlessly handles them by utilizing coarser objects as spatial anchors to progressively filter context, successfully resolving semantic ambiguity and pinpointing tiny targets that require rich hierarchical and spatial understanding.

To quantitatively evaluate this capability, we introduce a fine-grained CoR benchmark comprising 120 complex queries across four diverse scenes (30 queries

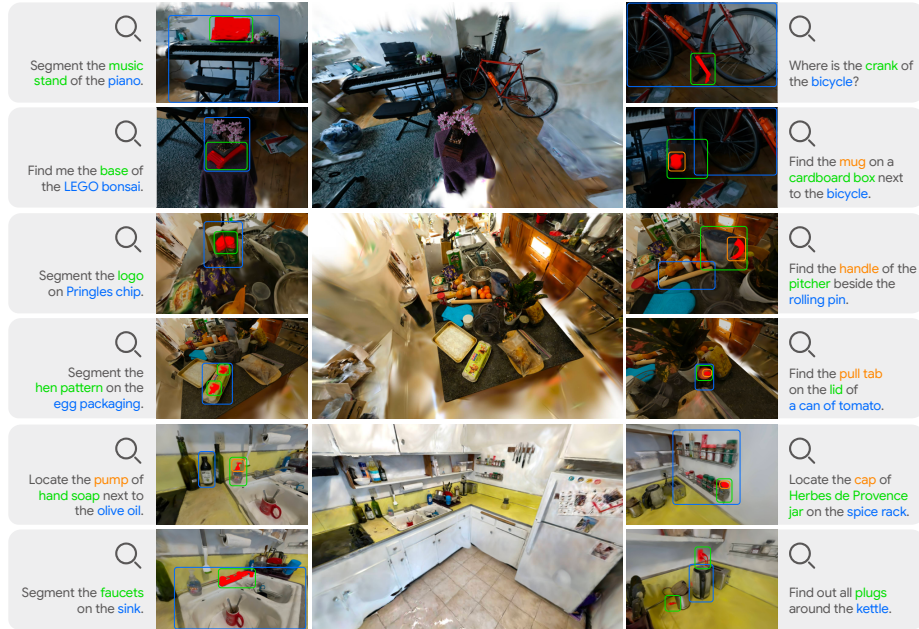


Fig. 7: Visualization of graph-based Chain-of-Retrieval. Colored words match bounding boxes, illustrating how coarser anchors guide fine-grained localization.

per scene, analogous to the qualitative examples in Fig. 7). Each query entails intricate part-whole and spatial relationships, primarily targeting fine-grained elements. As reported in Tab. 5, we compare LEGO against scene graph methods (THGS [9], BBQ [26]) and a strong baseline (LaGa). Note that LaGa is evaluated both in its native state and when augmented with our CoR strategy. BBQ and THGS construct scene graphs lacking semantic levels and consequently perform poorly. While adding a graph module to LaGa yields noticeable gains, its fixed, sparse SAM granularities (limited to 3 levels) suffer from multi-view inconsistency and fail to capture the accurate hierarchy essential for mixed relational querying. In contrast, LEGO maintains robust performance and achieves state-of-the-art results by unifying 3D-consistent semantic hierarchies with spatial relations.

4.2 Ablation Studies

We validate our core loss designs on the complex *Room* scene in Tab. 6. Omitting the scale-balancing re-weighting drops mAcc by 6.9% and mIoU by 3.7%, as gradients become dominated by massive entities, hindering the feature learning of fine-grained objects. Removing the auxiliary $\mathcal{L}_{pos}$ decreases mIoU by 1.3%; without this L_2 penalty, cosine similarity alone yields loose intra-cluster cohesion and ambiguous boundaries. Finally, discarding feature normalization ($\mathcal{L}_{3d}$,

Table 5: Quantitative evaluation of CoR queries (mIoU, %). We compare LEGO against scene graph baselines (BBQ, THGS) and a strong baseline LaGa, where LaGa is evaluated both natively and when augmented with our CoR strategy.

Method	<i>Teatime</i>	<i>Kitchen</i>	<i>Bonsai</i>	<i>Counter</i>	Overall
BBQ [26]	5.6	2.6	8.1	8.2	6.1
THGS [9]	9.9	5.0	10.3	12.9	9.5
LaGa [6]	5.8	5.2	12.0	10.4	8.3
LaGa w/ CoR	8.4	14.5	13.8	20.5	14.3
LEGO w/o CoR	25.2	10.0	30.2	25.6	22.7
LEGO(Ours)	52.2	44.9	57.6	51.4	51.6

Table 6: Ablation studies on the *Room* scene of the Mip-NeRF 360 dataset. We progressively analyze the impact of our scale-balancing re-weighting ($\mathcal{W}$), positive L_2 penalty ($\mathcal{L}_{pos}$), and feature normalization ($\mathcal{L}_{3d/2d}$).

Settings			Metrics	
Re-weighting ($\mathcal{W}$)	pos ($\mathcal{L}_{pos}$)	Norm ($\mathcal{L}_{3d/2d}$)	mAcc(%)	mIoU(%)
✓	✓	✓	93.1	67.3
	✓	✓	86.2	63.6
		✓	86.2	62.3
			82.8	60.7

$\mathcal{L}_{2d}$) further degrades mIoU by 1.6%. Without length regularization, the model exploits view-dependent shortcuts, causing feature decay during alpha-blending and resulting in inconsistent 3D semantic fields.

5 Conclusion

In this work, we introduced LEGO, a novel framework for multi-granular 3D scene decomposition and understanding. By shifting to intrinsic structural levels and utilizing a decoupled feature space, LEGO effectively prevents cross-level semantic entanglement and ensures consistent open-vocabulary supervision. Extensive experiments demonstrate state-of-the-art performance on both promptable and open-vocabulary benchmarks, uniquely excelling at extreme fine-grained part localization. Furthermore, by structuring these hierarchies into 3D scene graphs, LEGO enables LLM-guided Chain-of-Retrieval, successfully bridging the gap between flat 3D semantic fields and complex spatial reasoning.

Acknowledgements

This research was jointly supported by the NSFC Ph.D. Project (No. 424B2012) and NSFC General Project (No. 42571521).

References

1. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5470–5479 (2022)
2. Bhalgat, Y., Laina, I., Henriques, J.F., Zisserman, A., Vedaldi, A.: N2f2: Hierarchical scene understanding with nested neural feature fields. In: European Conference on Computer Vision. pp. 197–214. Springer (2024)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Segment any 3d gaussians. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1971–1979 (2025)
5. Cen, J., Fang, J., Zhou, Z., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Segment anything in 3d with radiance fields. *International Journal of Computer Vision* pp. 1–23 (2025)
6. Cen, J., Zhou, X., Fang, J., Wen, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Tackling view-dependent semantics in 3d language gaussian splatting. *arXiv preprint arXiv:2505.24746* (2025)
7. Cen, J., Zhou, Z., Fang, J., Shen, W., Xie, L., Jiang, D., Zhang, X., Tian, Q., et al.: Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems* **36**, 25971–25990 (2023)
8. Cheng, J., Zaech, J.N., Van Gool, L., Paudel, D.P.: Occam’s lgs: An efficient approach for language gaussian splatting. *arXiv preprint arXiv:2412.01807* (2024)
9. Dai, S., Qu, Y., Li, Z., Li, X., Zhang, S., Cao, L.: Training-free hierarchical scene understanding for gaussian splatting with superpoint graphs. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3673–3682 (2025)
10. Duisterhof, B.P., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In: 2025 International Conference on 3D Vision (3DV). pp. 1–10. IEEE (2025)
11. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European conference on computer vision. pp. 540–557. Springer (2022)
12. Goel, R., Sirikonda, D., Saini, S., Narayanan, P.: Interactive segmentation of radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4201–4211 (2023)
13. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
14. Hu, X., Wang, Y., Fan, L., Luo, C., Fan, J., Lei, Z., Li, Q., Peng, J., Zhang, Z.: Sagd: Boundary-enhanced segment anything in 3d gaussian via gaussian decomposition. *arXiv preprint arXiv:2401.17857* (2024)
15. Huang, T., Chen, R., Hu, D., Huang, F., Gong, M., Liu, T.: Openinsgaussian: Open-vocabulary instance gaussian segmentation with context-aware cross-view fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 6400–6409 (October 2025)

16. Jain, U., Mirzaei, A., Gilitschenski, I.: Gaussiancut: Interactive segmentation via graph cut for 3d gaussian splatting. *Advances in Neural Information Processing Systems* **37**, 89184–89212 (2024)
17. Ji, Y., Zhu, H., Tang, J., Liu, W., Zhang, Z., Tan, X., Xie, Y.: Fastlgs: Speeding up language embedded gaussians with feature grid mapping. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 3922–3930 (2025)
18. Jun-Seong, K., Kim, G., Yu-Ji, K., Wang, Y.C.F., Choe, J., Oh, T.H.: Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 14137–14146 (2025)
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
20. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 19729–19739 (2023)
21. Kim, C.M., Wu, M., Kerr, J., Goldberg, K., Tancik, M., Kanazawa, A.: Garfield: Group anything with radiance fields. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21530–21539 (2024)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4015–4026 (2023)
23. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546* (2022)
24. Li, H., Wu, Y., Meng, J., Gao, Q., Zhang, Z., Wang, R., Zhang, J.: Instance-gaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 14078–14088 (2025)
25. Li, W., Zhao, Y., Qin, M., Liu, Y., Cai, Y., Gan, C., Pfister, H.: Langsplatv2: High-dimensional 3d language gaussian splatting with 450+ fps. *arXiv preprint arXiv:2507.07136* (2025)
26. Linok, S., Zemskova, T., Ladanova, S., Titkov, R., Yudin, D., Monastyrny, M., Valenkov, A.: Beyond bare queries: Open-vocabulary object grounding with 3d scene graph. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 13582–13589. IEEE (2025)
27. Liu, K., Zhan, F., Zhang, J., Xu, M., Yu, Y., El Saddik, A., Theobalt, C., Xing, E., Lu, S.: Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems* **36**, 53433–53456 (2023)
28. Lyu, W., Li, X., Kundu, A., Tsai, Y.H., Yang, M.H.: Gaga: Group any gaussians via 3d-aware memory bank. *arXiv preprint arXiv:2404.07977* (2024)
29. McInnes, L., Healy, J., Astels, S., et al.: hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**(11), 205 (2017)
30. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)* (2019)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021)
32. Mirzaei, A., Aumentado-Armstrong, T., Derpanis, K.G., Kelly, J., Brubaker, M.A., Gilitschenski, I., Levinshtein, A.: Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20669–20679 (2023)

33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
34. Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., Chen, C., Wu, Z.: 3d vision-language gaussian splatting. arXiv preprint arXiv:2410.07577 (2024)
35. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 815–824 (2023)
36. Peng, Y., Wang, H., Liu, Y., Wen, C., Dong, Z., Yang, B.: Gags: Granularity-aware feature distillation for language gaussian splatting. arXiv preprint arXiv:2412.13654 (2024)
37. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024)
38. Qu, Y., Dai, S., Li, X., Lin, J., Cao, L., Zhang, S., Ji, R.: Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 5328–5337 (2024)
39. Ren, Z., Agarwala, A., Russell, B., Schwing, A.G., Wang, O.: Neural volumetric object selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6133–6142 (2022)
40. Shen, Q., Yang, X., Wang, X.: Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. In: European Conference on Computer Vision. pp. 456–472. Springer (2024)
41. Shen, Y., Fu, C., Chen, P., Zhang, M., Li, K., Sun, X., Wu, Y., Lin, S., Ji, R.: Aligning and prompting everything all at once for universal visual perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13193–13203 (2024)
42. Shi, J.C., Wang, M., Duan, H.B., Guan, S.H.: Language embedded 3d gaussians for open-vocabulary scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5333–5343 (2024)
43. Sun, W., Zhou, Y., Jiao, J., Li, Y.: Cags: Open-vocabulary 3d scene understanding with context-aware gaussian splatting. arXiv preprint arXiv:2504.11893 (2025)
44. Tan, X., Ji, Y., Zhu, H., Xie, Y.: Fmlgs: Fast multilevel language embedded gaussians for part-level interactive agents. arXiv preprint arXiv:2504.08581 (2025)
45. Wang, H., Liu, Y., Liu, Z., Wang, W., Dong, Z., Yang, B.: Vistadream: Sampling multiview consistent images for single-view scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26772–26782 (2025)
46. Wang, X., Yang, D., Gao, Y., Yue, Y., Yang, Y., Fu, M.: Gaussiangraph: 3d gaussian-based scene graph generation for open-world scene understanding. arXiv preprint arXiv:2503.04034 (2025)
47. Wang, Z., Su, Y., Li, C., Wang, D., Huang, Y., Li, X., Zhao, B.: Open-vocabulary octree-graph for 3d scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7037–7047 (2025)
48. Werby, A., Huang, C., Büchner, M., Valada, A., Burgard, W.: Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In: First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024 (2024)

49. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., et al.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *Advances in Neural Information Processing Systems* **37**, 19114–19138 (2024)
50. Yang, Z., Yang, B., Dong, W., Cao, C., Cui, L., Ma, Y., Cui, Z., Bao, H.: Instascene: Towards complete 3d instance decomposition and reconstruction from cluttered scenes. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7771–7781 (2025)
51. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: *European conference on computer vision*. pp. 162–179. Springer (2024)
52. Ying, H., Yin, Y., Zhang, J., Wang, F., Yu, T., Huang, R., Fang, L.: Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20612–20622 (2024)
53. Zhan, C., Zhang, Y., Wang, G., Wang, H.: Hi-lsplat: Hierarchical 3d language gaussian splatting. *arXiv preprint arXiv:2506.06822* (2025)
54. Zhang, J., Jiang, J., Chen, Y., Jiang, K., Liu, X.: Cob-gs: Clear object boundaries in 3dgs segmentation based on boundary-adaptive gaussian splitting. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 19335–19344 (2025)
55. Zhao, Y., Xu, W., Zheng, R., Qiao, P., Liu, C., Chen, J.: isegman: Interactive segment-and-manipulate 3d gaussians. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 661–670 (2025)
56. Zhu, R., Guo, J., Guo, X., Liu, Z., Hui, K.H., Yin, W., Chen, K., Chen, W., Ren, W., Liu, Y., et al.: Eps3d: End-to-end feed-forward 3d panoptic segmentation. *arXiv preprint arXiv:2606.08980* (2026)
57. Zhu, R., Hui, K.H., Liu, Z., Wu, Q., Tang, W., Qiu, S., Heng, P.A., Fu, C.W.: Cos3d: Collaborative open-vocabulary 3d segmentation. *arXiv preprint arXiv:2510.20238* (2025)
58. Zhu, R., Qiu, S., Liu, Z., Hui, K.H., Wu, Q., Heng, P.A., Fu, C.W.: Rethinking end-to-end 2d to 3d scene segmentation in gaussian splatting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3656–3665 (2025)
59. Zhu, R., Qiu, S., Wu, Q., Hui, K.H., Heng, P.A., Fu, C.W.: Pcf-lift: Panoptic lifting by probabilistic contrastive fusion. In: *European Conference on Computer Vision*. pp. 92–108. Springer (2024)
60. Zuo, X., Samangouei, P., Zhou, Y., Di, Y., Li, M.: Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *International Journal of Computer Vision* **133**(2), 611–627 (2025)