ReLaGS: Relational Language Gaussian Splatting

Yaxu Xie[1,2] * Abdalla Arafa[1,2] * Alireza Javanmardi[1] Christen Millerdurai[1]
Jia Cheng Hu[3] Shaoxiang Wang[1,2] Alain Pagani[1] Didier Stricker[1,2]
[1]German Research Center for Artificial Intelligence (DFKI)
[2]RPTU Kaiserslautern-Landau [3]University of Modena and Reggio Emilia

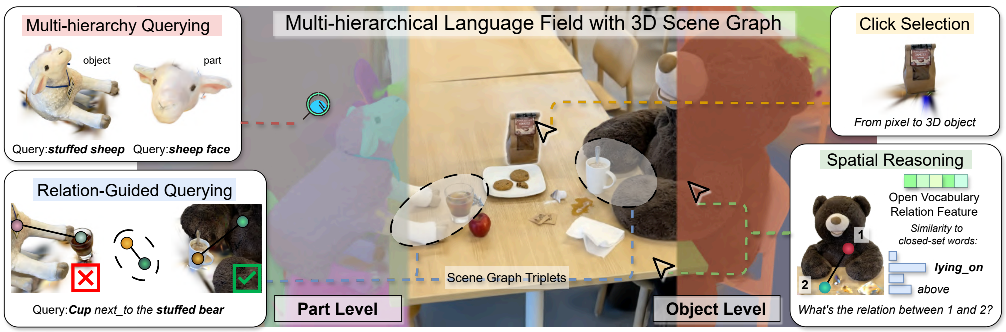


Figure 1. **Relational Language Gaussian Splatting**. We build a platform with multi-hierarchical language Gaussian field and open-vocabulary 3D scene graph, to support various tasks such as object selection via click, open vocabulary 3D object segmentation across semantic granularity, spatial relationship reasoning between objects and querying object with relation-guidance.

Abstract

Achieving unified 3D perception and reasoning across tasks such as segmentation, retrieval, and relation understanding remains challenging, as existing methods are either object-centric or rely on costly training for inter-object reasoning. We present a novel framework that constructs a hierarchical language-distilled Gaussian scene and its 3D semantic scene graph without scene-specific training. A Gaussian pruning mechanism refines scene geometry, while a robust multi-view language alignment strategy aggregates noisy 2D features into accurate 3D object embeddings. On top of this hierarchy, we build an open-vocabulary 3D scene graph with Vision Language-derived annotations and Graph Neural Network-based relational reasoning. Our approach enables efficient and scalable open-vocabulary 3D reasoning by jointly modeling hierarchical semantics and inter/intra-object relationships, validated across tasks including open-vocabulary segmentation, scene graph generation, and relation-guided retrieval. Project page: `https://dfki-av.github.io/ReLaGS/`

1. Introduction

Radiance-field representations such as NeRF [29] and Gaussian Splatting [15] have become core technologies for high-fidelity 3D reconstruction in VR/AR, robotics, and digital twins. Despite their impressive geometric and photometric accuracy, these representations lack meaningful scene semantics and therefore cannot support high-level reasoning. Recent *language field distillation* methods [8, 14, 16, 46, 51] address this gap by injecting vision–language priors into 3D radiance fields, enabling them to encode geometry, colors, and language embeddings distilled from 2D foundation models. This transforms purely geometric radiance fields into open-vocabulary 3D feature fields that can be queried directly in natural language, supporting tasks such as open-vocabulary segmentation [32, 46], language-guided navigation [4, 22], and scene editing [31, 40]. In such fields, users can issue queries such as *"find the wooden chair"* or *"highlight all green plants"* to retrieve semantically matched regions. However, despite these capabilities, existing methods remain fundamentally limited in semantic expressiveness. They fail in queries involving spatial relations or object parts—such as *"select the cup next to the laptop"* or *"highlight the keyboard of the laptop"*. These systems often segment the entire laptop instead of its keyboard or ignore spatial configurations altogether, as they operate with a single semantic granularity and without relational context. This reveals two key limitations of existing works: they remain **flat**, with single-level semantic abstraction and **isolated**, lacking inter-entity relationships. They describe what objects exist but not

*These authors contributed equally to this work how they relate, because their features remain confined within individual entities. Furthermore, these approaches lack a hierarchical organization of semantic abstraction, missing both finer part-level cues and higher-level contextual relationships essential for comprehensive and structured 3D scene understanding.

To overcome these limitations, recent efforts aim to embed relational context into radiance field [19, 41], bridging toward 3D scene graph formulations. RelationField [19] models relationships as ray pairs connecting object instances, learning spatial and functional dependencies between entities via multimodal language priors, but its volumetric rendering pipeline remains computationally heavy and memory intensive. SplatTalk [36] converts language-distilled Gaussians into semantic tokens for Large Language Model (LLM) reasoning, but tokenization and LoRA [11] fine-tuning remain costly and slow.

This puts us at a crossroads: we seek both multi-level scene representations capturing full *part–object–scene* hierarchy and relational 3D scene graphs that explain how entities interact. Yet existing methods achieve only one of these goals or rely on computationally expensive optimization. We therefore pursue both goals in a unified, optimization-free framework that organizes 3D scenes into coherent semantic hierarchies while also capturing their relational structure. To this end, we introduce a training-free pipeline that first constructs a language-grounded hierarchical Gaussian scene representation and then builds an explicit open-vocabulary 3D scene graph on top of it. Rather than assigning language and relation features to every Gaussian, we observe that semantics naturally emerge at coarser spatial granularity, while appearance details remain at the Gaussian level. Following this intuition, we cluster Gaussians in a bottom-up manner into sub-part, part, and object-level groups, regulated by multi-level masks. While segmentation-based methods [7, 46] achieve strong geometric grouping, they weakly handle language registration under inconsistent SAM masks and noisy CLIP [33] features. We address this challenge by introducing *Maximum Weight Pruning* and *Robust Outlier-Aware Feature Aggregation*, which jointly refine geometry and filter out unreliable language features, yielding consistent per-object language embeddings across multiple views and hierarchy levels.

On top of this hierarchy, we explicitly construct an open-vocabulary 3D scene graph that links entities through relational edges summarizing spatial and relationships. Our design naturally extends to intra-object graphs when finer-grained relational understanding is required. We propose two complementary variants for scene graph construction. The first lifts per-frame relational annotations obtained from LLM with Set-of-Mark (SoM) prompting [49] into 3D, providing high-quality but sparse relation edges. The second employs a pretrained graph transformer to infer relations directly from geometric and language features without additional costs and ensuring scalability, as we mainly evaluated in the experiments. Compared to RelationField, which requires hours of training and renders below 10 fps, our pipeline constructs a complete scene graph in under **15 minutes** and renders at over 200 fps. In summary, our framework unifies hierarchical scene construction and relational reasoning within 3D Gaussian fields, providing an explicit, scalable, and training-free foundation for structured 3D understanding. Our main contributions are as follows:

- We propose the first unified language distillation framework allowing for both hierarchical and relational reasoning on one single Gaussian field.
- We introduce Maximum Weight Gaussian pruning and robust outlier-aware feature aggregation methods, largely improving the geometric and language registration accuracy of the hierarchical Gaussian scene reconstruction pipeline.
- We provide two solutions to build explicit open-vocabulary 3D scene graphs: lifting 2D LLM-based annotations into 3D, and predicting relationships with a pretrained lightweight graph neural network.
- We demonstrate that our design achieves structured, open-vocabulary 3D scene understanding without scene-specific training, efficiently combining geometry, language, and relationships within a unified scene.

2. Related Work and Motivation

Language Field Distillation. Existing approaches for embedding language semantics into 3D Gaussian or radiance fields can be broadly categorized as *training-based* and *training-free*. Training-based methods [16, 30, 34, 51] incorporate vision–language supervision directly into the rendering and optimization loop of the radiance field. Although effective, per-scene training is inefficient in time and resource, and remains highly sensitive to view-inconsistent 2D features, making them impractical for large-scale deployment. To improve scalability, training-free and heuristic forward distillation methods [5, 14, 39, 47] lift 2D feature maps into 3D Gaussians through weighted or closed-form aggregation without gradient updates. Occam's LGS [5] formulates this as a MAP estimation problem solvable in closed form. Dr.Splat [14] adopts a similar scheme with top-k truncation for efficiency, and VALA [39] introduces a visibility-aware gating mechanism to suppress occluded contributions. Splat Feature Solver [47] further interprets feature lifting as a sparse linear inverse problem and proves that weighted aggregation methods [14, 21, 28] are bounded approximations to its least-squares solution. THGS [7] stands out among these by extending the *grouping-then-registration* strategy [24, 46] into a

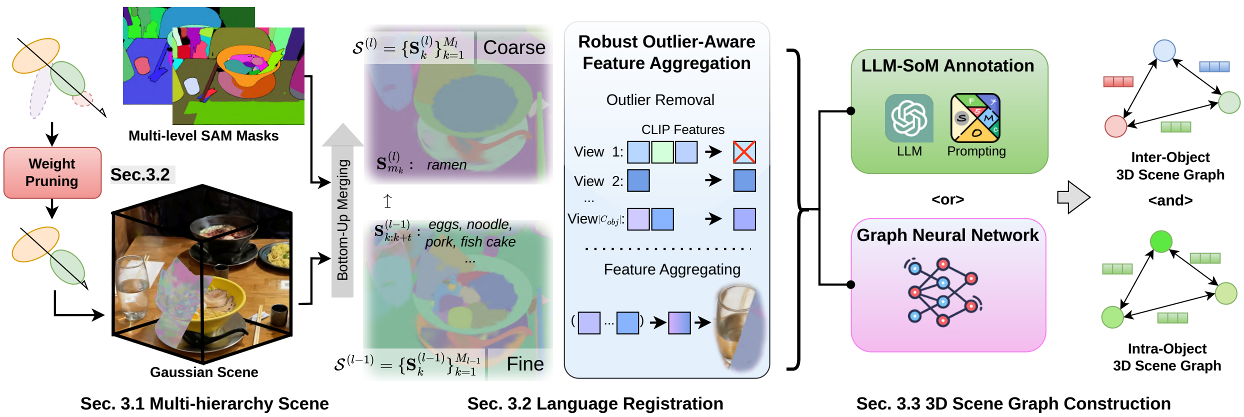


Figure 2. **ReLaGS Overview**. Given a reconstructed Gaussian scene, redundant primitives are first pruned to improve geometric accuracy. Heuristic clustering under multi-level SAM supervision then forms a hierarchical scene structure, where each cluster is assigned a CLIP-based language feature with outlier rejection. Finally, open-vocabulary inter- and intra-object scene graphs are obtained either by lifting LLM-derived relations for semantic diversity or by using a pretrained graph network for efficient offline inference.

training-free paradigm. It hierarchically merges Gaussians from superpoints to parts and objects, guided by multi-level SAM masks, forming a nested and interpretable structure that serves as the foundation for our scene construction.

3D Scene Graph [2, 38] is a structured graph representation that links objects, their spatial relationships, and semantic attributes in 3D scenes. Although many 3D scene graph generation works [27, 38, 42, 44, 45] are limited by closed-set semantic categories and highly rely on geometric reasoning in pre-segmented point cloud, recent advances have moved toward open-vocabulary 3D scene graph generation [18, 43], which integrates vision language models to associate 3D entities with rich semantics aligned to the language. ConceptGraphs [10], GaussianGraph [41], and RelationField [19] all aim to build open-vocabulary 3D scene graphs by combining vision–language features with geometry. ConceptGraphs depends on costly LLM inference and outputs text-based graphs, GaussianGraph requires scene-specific training and encodes relations implicitly through geometric heuristics, and RelationField learns relation fields via per-scene optimization, resulting in high computational cost and limited explicitness.

Our Motivation is to ***efficiently*** construct an ***explicit*** open-vocabulary 3D scene graph that connects the ***hierarchical*** Gaussian language field for structured ***causal reasoning*** in 3D. To ensure efficiency and explicitness, we adopt a training-free forward distillation framework [7] and employ a lightweight pretrained graph network to predict open-vocabulary relation embeddings from object language features and spatial cues. Alternatively, object-ID maps rendered from hierarchical scenes can be paired with SoM-prompted multimodal LLMs to annotate 2D relations, which are then lifted into 3D. Thus, inter-object relationships are represented explicitly as *object–predicate–subject* triplets, where both nodes and edges are discrete, language-grounded entities within the Gaussian scene. Because scene graph construction depends on accurate object segmentation and language registration, we further improve modules to ensure reliable relationship prediction. The hierarchical organization dictates how relationships are modeled across abstraction levels: inter-object graphs capture global spatial and functional relations, while intra-object graphs describe fine-grained part structures within each entity. Relations between parts of different objects (e.g., a *door handle* and a *table leg*) are generally meaningless and are instead summarized by their parent objects (e.g., *table next to door*), ensuring coherent and interpretable scene graphs. This hierarchical separation not only preserves clarity but also supports **causal reasoning** across levels—linking local part composition to global spatial context—thereby enabling structured understanding of how entities and their interactions collectively form a 3D scene.

3. Approach

Our method, **ReLaGS**, unifies hierarchical scene representation and relational reasoning within a single framework without scene-specific training. As illustrated in Fig. 2, given a reconstructed 3D Gaussian field, ReLaGS first constructs a multi-hierarchy representation that organizes the scene into nested levels of semantic abstraction, from fine-grained parts to complete objects. This hierarchical structure provides a compact and interpretable foundation for reasoning, language registration, and relationship prediction. We then explicitly build 3D scene graphs whose nodes correspond to hierarchical entities and whose edges encode semantic relations, either lifted from large language models or predicted by a pretrained graph neural network. We next formalize each stage of this pipeline, beginning with the hierarchical Gaussian representation.

3.1. Multi-Hierarchy Gaussian Representation

Let $\mathcal{G} = \{G_i\}_{i=1}^{N}$ denote the set of 3D Gaussian primitives, each encoding spatial and radiance attributes. We define a hierarchical representation with L abstraction levels:

$$\mathcal{S}^{(1)}, \dots, \mathcal{S}^{(L)}, \quad \mathcal{S}^{(l)} = \{\mathbf{S}_k^{(l)}\}_{k=1}^{M_l}, \quad k \in \{1, \dots, M_l\} \tag{1}$$

where $\mathcal{S}^{(l)}$ is the set of clusters at level l and M_l is the number of clusters at that level. Each Gaussian G_i is assigned to exactly one cluster at each level, and the clusters form a nested hierarchy:

$$\forall\, \mathbf{S}_k^{(l-1)} \in \mathcal{S}^{(l-1)}, \quad \mathbf{S}_k^{(l-1)} \subseteq \mathbf{S}_{m_k}^{(l)}, \quad \mathbf{S}_{m_k}^{(l)} \in \mathcal{S}^{(l)}. \tag{2}$$

Each cluster $\mathbf{S}_k^{(l)}$ represents a semantic entity at level l. For instance in Fig. 2 $\mathbf{S}_{m_k}^{(3)}$ correspond to a complete object (e.g., a *ramen*), while lower levels such as $\mathbf{S}_k^{(2)}$ capture finer-grained parts (e.g., the *eggs* as its ingredient). Each cluster carries an embedding $\mathbf{f}_k^{(l)} \in \mathbb{R}^d$, which encodes its language-aligned semantics. This induces a semantic hierarchy, $\mathcal{G} \rightarrow \mathcal{S}^{(1)} \rightarrow \cdots \rightarrow \mathcal{S}^{(L)}$, where higher levels represent coarse semantic granularity.

Rasterization and 2D-3D Tracing. To support tasks such as multi-hierarchy instance segmentation via point prompts on rendered images and lifting 2D scene graph annotations into 3D (see Sec. 3.3), we establish a consistent pixel-to-Gaussian correspondence through rasterization. The color $\mathbf{C}$ of each pixel (u, v) is obtained by volumetric α-blending along its viewing ray:

$$\mathbf{C}_{(u,v)} = \sum_{i=1}^{N} \mathbf{c}_i\, \alpha_i\, T_i, \quad T_i = \prod_{j<i} (1 - \alpha_j), \tag{3}$$

where $\mathbf{c}_i$ and α_i denote the color and opacity of Gaussian G_i, and T_i represents the accumulated transmittance of all preceding Gaussians along the ray. We compute the contribution weight $w_i = \alpha_i T_i$ for each Gaussian and define the dominant Gaussian for pixel (u, v) as:

$$G^*_{(u,v)} = \arg\max_i\, w_i. \tag{4}$$

This dominant Gaussian provides a reliable 2D-3D correspondence, which naturally extends to map a pixel to cluster $G^*_{(u,v)} \in \mathbf{S}_k^{(l)}$.

3.2. Improving Construction and Language Lifting of Hierarchical Scene

To construct the multi-hierarchy scene representation described in Sec. 3.1, we adopt THGS [7], which organizes the Gaussian field without gradient-based optimization. The generated field $\mathcal{G}$ is partitioned into geometrically coherent superpoints via Cut Pursuit [20] and progressively merged into higher-level clusters $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \dots, \mathcal{S}^{(L)}$ by balancing intra-cluster compactness and inter-cluster boundary saliency, guided by SAM mask priors [17]. Further details can be found in THGS [7]—Secs. 3.2–3.4.

Since accurate segmentation and language feature of objects are both critical for scene graph prediction and reconstruction, we augment THGS with two novel complementary steps that enhance the multi-hierarchy scene representation by refining both geometric and semantic quality, yielding more accurate object–part boundaries and improved language-aligned object representations. First, Maximum Weight Pruning (MWP) eliminates Gaussians with negligible visual impact across all training views, ensuring that only geometrically relevant components are retained. Second, Robust Outlier-Aware Feature Aggregation (ROFA) enhances the reliability of object-level semantics by filtering inconsistent CLIP features before aggregation. Together, these steps yield a compact yet semantically consistent representation that serves as a strong foundation for downstream reasoning tasks.

Maximum Weight Pruning. While THGS constructs its hierarchy directly from the raw Gaussian field, we observe that the unfiltered representation often contains numerous *floaters*—Gaussians that contribute negligibly to the rendered views. These floaters, typically found near object boundaries or in occluded regions, introduce spurious pixel-aligned language features and lead to fragmented or noisy object-level clusters. Such artifacts become increasingly detrimental in multi-level hierarchical grouping, where boundary precision is crucial for maintaining semantic consistency across abstraction levels (see Fig. 3 (a)). To address this limitation, we introduce a *maximum weight pruning* strategy that removes inconsistent or redundant Gaussians prior to hierarchical construction, thereby improving boundary integrity and the overall stability of the semantic hierarchy.

Building on Eq. 4, we compute the maximum contribution of each Gaussian across all camera views as:

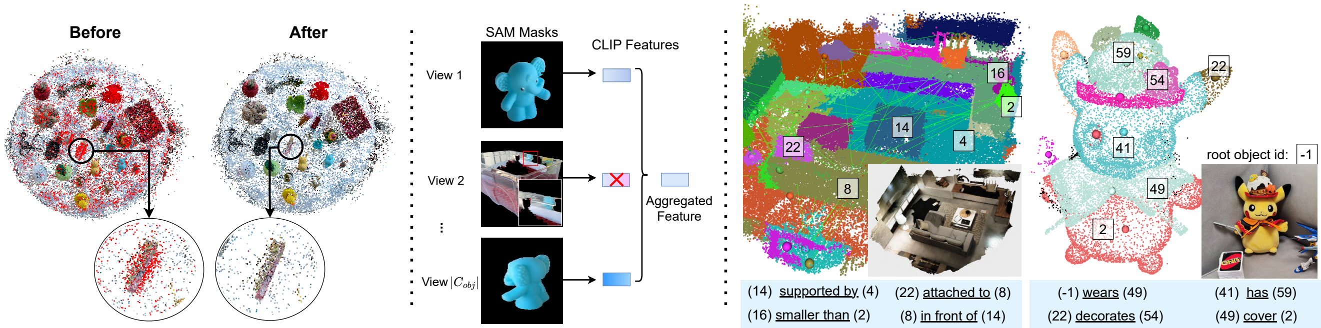
$$\omega_i^{\max} = \max_{c \in \mathcal{C}, p \in \mathcal{P}_c} w_{i,p}^{(c)}, \tag{5}$$

where $w_{i,p}^{(c)}$ denotes the contribution weight of Gaussian i for pixel p in camera view c. Here, $\mathcal{C}$ represents the set of training camera views, and $\mathcal{P}_c$ denotes the set of pixels in camera view c. Gaussians whose maximum contribution weights falls below a small threshold $\tau_{contrib}$ are pruned from the scene:

$$\mathcal{G}' = \{\, G_i \in \mathcal{G} \mid \omega_i^{\max} > \tau_{contrib} \,\}. \tag{6}$$

This filtering step removes geometrically inconsistent or redundant Gaussians arising from occlusion or low opacity regions, while preserving the overall rendering quality. As demonstrated in our ablation study (Sec. 4.4), the proposed pruning substantially improves geometric consistency and enhances the reliability of subsequent object-level clustering.

Robust Outlier-Aware Feature Aggregation. Multi-view language registration often suffers from inconsistent CLIP



(a) Maximum Weight Pruning (b) Robust Outlier-Aware Feature Aggregation (c) Inter-object 3D scene graph (d) Intra-object 3D scene graph

Figure 3. Illustration of proposed two improvement methods for hierarchical scene construction and two example scene graphs. **(a):** Low contribution Gaussian points (red) are removed to improve scene geometry. **(b):** Outlier features (e.g., due to occluded or inconsistent masks) are filtered before aggregation, producing a more coherent and consistent embedding for the target. **(c):** The spatial relationships are predicted by our GNN. **(d):** The more semantic-enriched relationship lifted with LLM, the root object is marked as -1.

features caused by erroneous SAM masks or extreme viewpoints. Directly averaging features across views, as done in THGS makes the object embedding sensitive to these outliers (see Fig.3 (b). Given an object visible in $\mathcal{C}_{obj}$ views with corresponding CLIP features $\{\mathbf{f}_i\}_{i=1}^{|\mathcal{C}_{obj}|}$, we first measure the semantic consistency of each feature by computing its mean cosine similarity to all others:

$$s_i = \frac{1}{|\mathcal{C}_{obj}|-1}\sum_{j\neq i}\frac{\mathbf{f}_i\cdot\mathbf{f}_j}{\|\mathbf{f}_i\|\|\mathbf{f}_j\|}. \tag{7}$$

We then perform Z-score normalization over the similarity scores:

$$z_i = \frac{s_i-\mu_s}{\sigma_s}, \tag{8}$$

where μ_s and σ_s denote the mean and standard deviation of $\{s_i\}$. Features with $z_i < -\tau_{lang}$ are treated as low-similarity outliers and removed. Here, τ_{lang} is a Z-score threshold that controls the sensitivity of outlier detection. The final object-level embedding is obtained by averaging the remaining filtered features. Our Robust Outlier-Aware Feature Aggregation (ROFA) suppresses multi-view inconsistencies and yields more reliable language-aligned object representations. As shown in our ablation study (Sec. 4.4), this filtering leads to more stable and semantically consistent object embeddings.

3.3. Relation Lifting and Prediction

Our hierarchical Gaussian scene enables explicit 3D scene graphs at different abstraction levels, as illustrated in Fig. 3 (c) and (d), a novel capability that brings structured, interpretable, and multi-level reasoning to Gaussian fields. At the top level L (object level), we define an *inter-object* graph capturing semantic relationship:

$$\mathcal{H}_{\text{3D}}^{(L)} = (\mathcal{V}_{\text{obj}}^{(L)}, \mathcal{E}_{\text{inter}}^{(L)}, F_V^{(L)}, F_E^{(L)}), \tag{9}$$

where $\mathcal{V}_{\text{obj}}^{(L)}$ and $\mathcal{E}_{\text{inter}}^{(L)}$ denote object nodes and their pairwise relations, with corresponding node and edge features $F_V^{(L)}$ and $F_E^{(L)}$. Each node $v_k^{(L)}$ inherits its language-aligned feature $\mathbf{f}_k^{(L)}$, while each edge $(v_i, v_j) \in \mathcal{E}_{\text{inter}}^{(L)}$ carries a relation embedding $\mathbf{r}_{ij}^{(L)}$ encoded by Jina [35]. At lower hierarchy levels $l < L$, we form *intra-object* graphs to describe relationships among parts within the same object:

$$\begin{aligned}\mathcal{H}_{\text{3D,intra}}^{(l)} &= (\mathcal{V}_{\text{part}}^{(l)}, \mathcal{E}_{\text{intra}}^{(l)}, F_V^{(l)}, F_E^{(l)}),\\ \mathcal{E}_{\text{intra}}^{(l)} &= \{(v_i, v_j)\,|\, v_i, v_j \in S_m^{(L)}\},\end{aligned} \tag{10}$$

where $S_m^{(L)}$ is the Gaussian cluster of object m. The inter-object graph $\mathcal{H}_{\text{3D}}^{(L)}$ captures global scene relations, while the intra-object graphs $\mathcal{H}_{\text{3D,intra}}^{(l)}$ encode local part composition and affordance structure. We next describe how both relation types can be either lifted using LLM-based annotation or predicted using pretrained GNN reasoning.

3D Consistent Relation Lifting from LLM Annotation. To obtain open-vocabulary relational annotations, we follow the SoM–LLM paradigm also used in RelationField [19], but differ in how 2D object masks are acquired. In [19], per-frame instance masks is generated with SAM [17], which often suffer from view inconsistency and mixed semantic granularity. In contrast, we leverage our hierarchical Gaussian scene and Eq. 4 to render per-view Gaussian ID maps, trace them to their corresponding clusters, and thus obtain view-consistent cluster ID maps. Each cluster mask is then overlaid with numeric marks with the SoM prompting strategy [49], enabling a MLLM such as GPT-4V to infer relationships between the marked object pairs. For every pair of objects, the LLM outputs textual predicates $\langle s, p, o\rangle$ (subject, predicate, and object) as 2D relational annotations. Because each mask corresponds to a 3D-consistent cluster ID, lifting these annotations into 3D is straightforward: we iterate over all annotated relations and assign them to the corresponding cluster pairs. After processing all frames, we pick the top-k_p frequent predicates of an edge from all gathered predicates, encode the texts using Jina [35] and average to form the lifted relational embedding $\mathbf{f}_{ij}$.

Relation Prediction with GNN. As explained in Sec. 1, the lifted 3D scene graph has limited coverage of all plausible relationship in the scene. On the other hand, the latent space of open vocabulary spatial relationships is less diverse than the latent space describing the texture, shape, material and use of objects. This motivates us to distill the open vocabulary relationship reasoning into a light weight pretrained GNN. We first construct a neighboring graph $\mathcal{H}'$ by connecting object nodes within a fixed distance threshold. On top of this graph, we use the residual graph neural network $\mathcal{F}_\theta$ to predict relation embeddings:

$$\hat{\mathbf{f}}_{ij} = \mathbf{f}'_{ij} + \mathcal{F}_\theta(\mathbf{f}_v^{\text{src}}, \mathbf{f}_v^{\text{dst}}, \mathbf{f}'_{ij}), \tag{11}$$

where $\mathbf{f}_v^{\text{src}}$ and $\mathbf{f}_v^{\text{dst}}$ denote the language–geometry fused node embeddings of the source and destination objects (see Appendix 6 for more details). The network outputs open-vocabulary edge features $\hat{\mathbf{f}}_{ij} \in \mathbb{R}^{d_r}$ with the same embedding space as Jina-Embedding-V3, enabling direct cosine-similarity comparison with textual predicates. We pretrain $\mathcal{F}_\theta$ on the 3RScan dataset [37] following the setup of Open3DSG [18], using a contrastive learning objective between predicted and ground-truth relation embeddings to align features across modalities. Because the modality gap between language-lifted Gaussians and point-cloud–image features is small, the pretrained model generalizes well to our Gaussian domain and can be applied directly for inference without fine-tuning, as shown in Sec. 4.2.

3.4. Applications: Hierarchical and Relational Reasoning in 3D

Our hierarchical scene representation and explicit 3D scene graph jointly enable structured reasoning within Gaussian fields. The hierarchy supports *compositional reasoning* by capturing how parts form objects (additional analysis of hierarchical queries at object and part levels is provided in Appendix 8), while the scene graph facilitates *relational reasoning* by modeling how objects interact. To the best of our knowledge, we are the first to unify these two forms of reasoning within Gaussian fields, providing a causal understanding of the scene that links "what exists," "how it is composed," and "how it relates to each other." We demonstrate this capability through multi-hierarchy querying and triplet-based relational querying.

Language-based multi-hierarchy querying. In practical situations involving natural language, people may describe either an entire object (*"ramen"*) or one of its parts (*"noodle"*), and may also refer to multiple similar instances simultaneously. To handle these ambiguities, our goal is twofold: (1) automatically determine whether the query best matches a root-level object or a finer leaf-level cluster, and (2) identify all relevant matches when multiple clusters exhibit comparable similarity. We design a tree-searching method to solve the querying problem analytically, following a simple intuition: a text referring to the sub-object part (leaf cluster) has higher similarity to the leaf language feature than the root. We compute the cosine similarity of the CLIP embedding of a text query $\mathtt{t}$ with per-cluster features $\mathbf{f}_k^{(l)}$ across hierarchy levels. Unlike the flat top-k retrieval in [7, 13], we search over the nested cluster structure $\mathcal{S}^{(1:L)}$ to adaptively match query granularity. Starting from root-level candidates, the search descends to child clusters whenever they exhibit higher similarity, distinguishing whether the query refers to a whole object or one of its parts. We further detect the largest drop in the similarity curve to automatically select multiple valid matches when present. The resulting union of Gaussians forms an adaptive segmentation mask that unifies coarse object localization and fine part-level discovery within a single query, see Algm. 1 in Appendix.

Language-Triplet querying on 3D Scene Graph. Beyond single-entity segmentation, we extend our querying framework to handle relationship-based queries of the form $\langle s, p, o \rangle$. Given such a query, our goal is to retrieve the 3D Gaussians corresponding to the subject entities that satisfy the specified relationship with the object. To achieve this, we operate on the 3D scene graph constructed atop our multi-hierarchical scene representation. The search for potential subjects and objects follows the same multi-hierarchy querying procedure described above, ensuring that both fine-grained clusters and high-level object groups are considered as valid candidates. After identifying candidate subject–object pairs, we evaluate their relationships by comparing the corresponding edge embeddings to the predicate embedding from the query. Each valid pair is then ranked using three complementary similarity measures: (1) subject–text alignment, (2) object–text alignment, and (3) predicate–relation alignment in the Jina embedding space. The Gaussians associated with the highest-ranked subject candidates are finally returned as the 3D regions most consistent with the queried relationship.

4. Experiments

In this section, we evaluate our method from two aspects: first, 3D scene graphs including its prediction in Sec. 4.1 and the extended application, relationship-guided 3D object segmentation using the scene graph in Sec. 4.2, and second, the open vocabulary object segmentation and 3D semantic segmentation in Sec. 4.3. The best and second-best results of all experiments are marked in green and blue, respectively. We also provide ablation studies on our proposed different modules in Sec. 4.4. Implementation details are given in Appendix 7.

4.1. 3D Scene Graph Prediction

We evaluate our method on the RIO10 subset of 3DSSG [37] for 3D scene graph prediction. This dataset provides semantic 3D scene graphs for pre-segmented 3D point clouds with 160 object classes and 27 relationship categories. We compare against recent open-vocabulary 3D scene graph approaches, including Open3DSG [18], which operates directly on pre-segmented point clouds, ConceptGraph [10], which incrementally reconstructs scenes and scene graphs from RGB-D sequences, and RelationField [19]. We group them into scene-specific and scene-agnostic methods in Tab. 1, the same in Sec. 4.2. Quantitative results are reported with Recall@K for object, and relationship prediction following the protocols of [26, 48]. Despite ground-truth annotation in [37] exhibits uneven semantic granularity across scenes (e.g., carpets are merged with floors, while doors and door frames labeled separately) and sometime misaligned with our hierarchical representation, our method with GNN still achieves better Recall on relationship prediction than others, with an improvement of 0.3 R@3 and 0.5 R@5 over RelationField, and exhibits only a 0.1 worse object Recall. VLM-based method with 2D images as input performs poorly, due to the lack of spatial reasoning capability. Our method is also 4.7× faster and 7.6× more memory efficient than RelationField, see Tab. 7.

Table 1. **Results of 3D scene graph prediction on 3DSSG [37].** "Scene agnostic" denotes methods without per-scene training.

Method	Object R@5	Object R@10	Predicate R@3	Predicate R@5	Scene agno.
ConceptGraphs [10]	0.37	0.46	0.74	0.79	✗
RelationField [19]	0.69	0.80	0.76	0.82	✗
Ours (VLM)	0.68	0.79	0.10	0.35	✗
Open3DSG [18]	0.56	0.61	0.58	0.65	✓
Ours (pred.)	0.68	0.79	0.79	0.87	✓

4.2. Evaluation on Open Vocabulary Instance Segmentation with Relationship Guidance

We evaluate our method on **ScanNet++** [50] dataset with the benchmark provided by RelationField [19]. The task aims to segment the *subject* entity in 3D given an open-vocabulary relationship query represented as a triplet $\langle s, p, o \rangle$. We disable both densification and our pruning as suggested in OpenGaussian [46], to guarantee bijection between points in the ground truth point cloud and our Gaussian scene. So that we can compute 3D mean-IoU between segmented object in Gaussian scene and the ground truth point cloud.

Among all language field distillation methods [8, 16, 19, 30] we compared, only RelationField [19] and our approach are using scene graph for relationship-based queries searching. Other methods are given a single query concatenated from the triplet text. We use predicted 3D inter-object scene graph from our GNN, to investigate the cross-dataset generalization of our trained network. As shown in Tab. 2, our method achieves the highest performance among all approaches, surpassing both training-based and training-free baselines. In particular, our approach attains an mIoU of 0.56, outperforming the closest

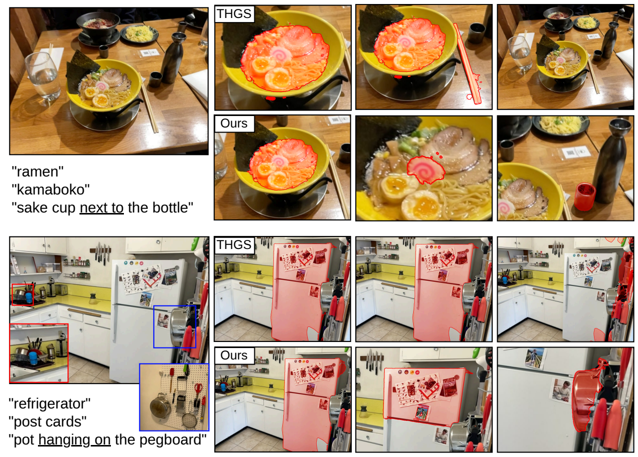


Figure 4. **Qualitative results of open vocabulary object segmentation.** We show results on LERF dataset for segmentation mask on 2D view. With multi-hierarchy querying search and 3D scene graph for relation guidance, our method shows strong improvement against THGS.

training-based competitor RelationField (0.53), despite requiring no additional training or LLM annotations. This improvement is attributed to our **multi-hierarchical scene representation** and **graph-based relational reasoning**, which together enable effective inference of spatial relationships purely from the 3D structure (see Fig. 4). Removing the multi-hierarchy (w/o MH) degrades performance, underscoring the role of hierarchical search in resolving complex spatial relationships.

Table 2. **Relationship-Guided 3D Instance Segmentation on ScanNet++ [50].** Our method achieves the highest mean IoU, even operating only on single hierarchy (w/o MH).

Method	mIoU	Scene agno.
Lerf [16]	0.25	✗
OpenNeRF [8]	0.45	✗
LangSplat [30]	0.49	✗
RelationField [19]	0.53	✗
THGS [7]	0.29	✓
Ours w/o MH	0.54	✓
Ours	0.56	✓

4.3. Open Vocabulary Object Querying

We evaluate our approach on two open-vocabulary segmentation benchmarks: LeRF-OVS [16] for object segmentation and ScanNet [6] for semantic segmentation. Tab. 3 and 4 report quantitative comparisons against recent methods, with training-free method specified (T-F). Our method achieves state-of-the-art performance on both datasets, outperforming both training-based and training-free approaches. On LeRF-OVS, we further analyze performance on hierarchical queries by separating object-level and part-level queries; detailed results are reported in Appendix 8. The most notable gains are observed in the *Figurines* and *Teatime* scenes, where the proposed feature aggregation and geometry refinement yield stronger semantic alignment under cluttered and occluded setups. We also provide qualitative results in Fig. 4. On ScanNet, our method achieves a slight improvement over This moderate gain is mainly attributed to the experiment setup, in which computing 3D-mIoU requires fixed Gaussian primitives number thus excluding our *Maximum Weight Pruning* component. Detailed analysis using a revised evaluation protocol, including densified Gaussian reconstruction and nearest-point assignment, is provided in Appendix 9.

Table 3. **Open-vocabulary segmentation on LERF-OVS [16].** Our method achieves the highest mean IoU among both training and training-free approaches.

Method	LERF-OVS mIoU (%)				Mean	T-F
	Fig.	Ramen	Teatime	Waldo		
LangSplatV2 [23]	56.4	51.4	72.2	59.1	59.9	✗
LAGA [3]	64.1	55.6	70.9	65.6	64.0	✗
THGS [7]	57.3	43.5	68.3	50.7	54.9	✓
Occam's [5]	58.6	51.0	70.2	65.3	61.3	✓
VALA [39]	59.9	51.5	70.2	65.1	61.7	✓
Ours	64.7	51.2	81.0	60.6	64.4	✓

Table 4. **Open-vocabulary semantic segmentation on ScanNet [6].** Our method achieves the highest mean performance among both training and training-free approaches on 15 and 10 classes subset.

Methods	19 cls		15 cls		10 cls		T-F
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	
LangSplat [30]	3.78	9.11	5.35	13.20	8.40	22.06	✗
OpenGau. [46]	24.73	41.54	30.13	48.25	38.29	55.19	✗
LAGA [3]	32.50	49.10	35.50	53.50	42.60	63.20	✗
THGS [7]	34.39	50.74	39.61	57.07	46.38	64.74	✓
Occam's [5]	31.93	48.93	34.25	53.71	45.16	64.39	✓
VALA [39]	32.11	50.05	35.10	54.77	46.21	65.61	✓
Ours	32.35	44.52	40.04	60.59	47.17	66.08	✓

4.4. Ablation Study

We evaluate the impact of our proposed components on open-vocabulary 2D segmentation (Tab. 5). Maximum Weight Pruning yields the largest improvement, underscoring the importance of refining the scene geometry by removing low-contribution Gaussians that otherwise introduce structural noise. Robust Outlier-Aware Feature Aggregation provides additional gains by filtering inconsistent language features across views, improving the semantic coherence of the aggregated embeddings. Notably, its effect is most pronounced on the *Figurines* and *Ramen* scenes of LeRF-OVS, where dense object arrangements and frequent occlusions (and, in *Ramen*, the presence of transparent surfaces) tend to corrupt the averaged semantic features.

To further analyze the sensitivity of the outlier detection threshold τ_{lang}, we report results for varying values in Tab. 6. The IoU initially increases with larger τ_{lang} as more consistent features are retained, but declines once the threshold becomes too permissive and introduces noise.

A runtime and memory consumption analysis is given in Tab. 7, in which we decompose the resource usage of our method in three stages, showing that our method can lift structure semantic representation to a Gaussian scene with less than 25% memory overhead. More ablation studies are given in the appendix about the graph neural network design, the pruning's effect on rendering quality, and the effect of our pruning and language lifting method on the final scene graph prediction.

Table 5. **Ablation** on LeRF-OVS dataset for open-vocabulary segmentation. M. is Maximum Weight Pruning and R. is Robust Outlier-Aware Feature Aggregation .

Variant	Figurines	Ramen	Teatime	Kitchen	Mean
Base	52.05	47.19	76.77	47.5	55.88
Base+M.	59.16	47.41	80.98	60.59	62.04
Base+M.+R.(Full)	64.69	51.15	80.98	60.6	64.36

Table 6. **Ablation** on different values of τ_{lang} on LeRF-OVS.

τ_{lang}	Figurines	Ramen	Teatime	Waldo Kitchen	Mean
1	61.85	43.3	76.02	63.37	60.78
2	63.39	51.07	73.8	60.62	62.22
3	64.69	51.15	80.98	60.60	64.36
4	63.3	47.4	80.98	60.60	63.07

Table 7. **Runtime and resource comparison** against RelationField [19] on 3DSSG scenes. Our method is 4.7× faster and 7.6× more memory efficient and a maximum of 7.5GB of GPU memory.

Method	Time (min) / Storage (MB) / GPU Memory (GB)			
	Scene Rec.	Language Distill.	Scene Graph	Overall
RF [19]	60/500/32 (inseparable)			60/500/32
Ours	11/52.2/3.2	1.5/10.6/7.5	0.1/2.2/3.0	12.6/65/7.5

5. Conclusion

We introduced ReLaGS, the first framework to unify multi-hierarchical 3D Gaussian fields and open-vocabulary 3D scene graphs within a single language-grounded representation. Through *Maximum Weight Pruning* and *Robust Outlier-Aware Feature Aggregation*, ReLaGS improves geometric fidelity, multi-view language registration, and the overall robustness of the scene representation. Building upon this unified scene representation, our explicit scene graph formulation enables scalable, training-free relational reasoning at far lower computational and memory costs than prior methods. ReLaGS achieves state-of-the-art performance across open-vocabulary segmentation, 3D scene graph prediction, and relationship-guided object segmentation, advancing open-vocabulary 3D understanding and paving the way for causal and compositional reasoning in complex real-world scenes. We hope our framework enables the community to efficiently generate large-scale, semantically enriched 3D radiance fields at low computational and resource cost.

Acknowledgements

This work has been partially funded by the EU projects dAIEDGE (Grant Agreement Number: 101120726) and LUMINOUS (Grant Agreement Number: 101135724).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5664–5673, 2019. 3

[3] Jiazhong Cen, Xudong Zhou, Jiemin Fang, Changsong Wen, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Tackling view-dependent semantics in 3d language gaussian splatting. In *ICML*, 2025. 8, 5

[4] Timothy Chen, Ola Shorinwa, Joseph Bruno, Aiden Swann, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager. Splat-nav: Safe real-time robot navigation in gaussian splatting maps. *IEEE Transactions on Robotics*, 2025. 1

[5] Jiahuan Cheng, Jan-Nico Zaech, Luc Van Gool, and Danda Pani Paudel. Occam's lgs: A simple approach for language gaussian splatting. 2024. 2, 8, 4, 5

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 7, 8, 5

[7] Shaohui Dai, Yansong Qu, Zheyan Li, Xinyang Li, Shengchuan Zhang, and Liujuan Cao. Training-free hierarchical scene understanding for gaussian splatting with superpoint graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3673–3682, 2025. 2, 3, 4, 6, 7, 8, 5

[8] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, Marc Pollefeys, and Federico Tombari. Opennerf: open set 3d neural scene segmentation with pixel-wise features and rendered novel views. *arXiv preprint arXiv:2404.03650*, 2024. 1, 7

[9] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pages 540–557. Springer, 2022. 2

[10] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 3, 7

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[12] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 2

[13] Minchao Jiang, Shunyu Jia, Jiaming Gu, Xiaoyuan Lu, Guangming Zhu, Anqi Dong, and Liang Zhang. Votesplat: Hough voting gaussian splatting for 3d scene understanding. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 6

[14] Kim Jun-Seong, GeonU Kim, Kim Yu-Ji, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14137–14146, 2025. 1, 2

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[16] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19729–19739, 2023. 1, 2, 7, 8, 5

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 4, 5

[18] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2024. 3, 6, 7, 4, 5

[19] Sebastian Koch, Johanna Wald, Mirco Colosi, Narunas Vaskevicius, Pedro Hermosilla, Federico Tombari, and Timo Ropinski. Relationfield: Relate anything in radiance fields. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21706–21716, 2025. 2, 3, 5, 7, 8, 4

[20] Loic Landrieu and Guillaume Obozinski. Cut pursuit: Fast algorithms to learn piecewise constant functions on general weighted graphs. *SIAM Journal on Imaging Sciences*, 10(4): 1724–1766, 2017. 4

[21] Hyunjoon Lee, Joonkyu Min, and Jaesik Park. Cf3: Compact and fast 3d feature fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27906–27916, 2025. 2

[22] Xiaohan Lei, Min Wang, Wengang Zhou, and Houqiang Li. Gaussnav: Gaussian splatting for visual navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[23] Wanhua Li, Yujie Zhao, Minghan Qin, Yang Liu, Yuanhao Cai, Chuang Gan, and Hanspeter Pfister. Langsplatv2:

High-dimensional 3d language gaussian splatting with 450+ fps. 2025. 8, 5

[24] Siyun Liang, Sen Wang, Kunyi Li, Michael Niemeyer, Stefano Gasperini, Nassir Navab, and Federico Tombari. Supergseg: Open-vocabulary 3d segmentation with structured super-gaussians. *arXiv preprint arXiv:2412.10231*, 2024. 2

[25] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023. 4, 5

[26] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 7, 5

[27] Changsheng Lv, Mengshi Qi, Xia Li, Zhengyuan Yang, and Huadong Ma. Sgformer: Semantic graph transformer for point cloud-based 3d scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4035–4043, 2024. 3, 1

[28] Juliette Marrie, Romain Ménégaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7440–7450, 2025. 2

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[30] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 2, 7, 8

[31] Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Language-driven physics-based scene synthesis and editing via feature splatting. In *European Conference on Computer Vision*, pages 368–383. Springer, 2024. 1

[32] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 5328–5337, 2024. 1

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[34] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024. 2

[35] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*, 2024. 5, 2

[36] Anh Thai, Songyou Peng, Kyle Genova, Leonidas Guibas, and Thomas Funkhouser. Splattalk: 3d vqa with gaussian splatting. *arXiv preprint arXiv:2503.06271*, 2025. 2

[37] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 6, 7, 3, 5

[38] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 3

[39] Sen Wang, Kunyi Li, Siyun Liang, Elena Alegret, Jing Ma, Nassir Navab, and Stefano Gasperini. Visibility-aware language aggregation for open-vocabulary segmentation in 3d gaussian splatting. In *ICCV 2025 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild*, 2025. 2, 8

[40] Shaoxiang Wang, Shihong Zhang, Christen Millerdurai, Rüdiger Westermann, Didier Stricker, and Alain Pagani. Inpaint360gs: Efficient object-aware 3d inpainting via gaussian splatting for 360deg scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 117–127, 2026. 1

[41] Xihan Wang, Dianyi Yang, Yu Gao, Yufeng Yue, Yi Yang, and Mengyin Fu. Gaussiangraph: 3d gaussian-based scene graph generation for open-world scene understanding. 2025. 2, 3

[42] Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, and Lu Sheng. Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21560–21569, 2023. 3

[43] Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14158–14168, 2025. 3

[44] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 3

[45] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Incremental 3d semantic scene graph prediction from rgb sequences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5064–5074, 2023. 3

[46] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *Advances in Neural Information Processing Systems*, 37:19114–19138, 2024. 1, 2, 7, 8, 5

[47] Butian Xiong, Rong Liu, Kenneth Xu, Meida Chen, and Andrew Feng. Splat feature solver. *arXiv preprint arXiv:2508.12216*, 2025. 2

[48] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 7, 5

[49] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 2, 5

[50] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 7, 5

[51] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 1, 2