GAGS: Granularity-Aware Feature Distillation for Language Gaussian Splatting

Yuning Peng[1]*, Haiping Wang[1]*, Yuan Liu[2], Chenglu Wen[3], Zhen Dong[1]†, Bisheng Yang[1]

[1]Wuhan University
[2]Hong Kong University of Science and Technology
[3]Xiamen University
yuningpeng@whu.edu.cn, hpwang@whu.edu.cn, yuanly@connect.hku.hk, clwen@xmu.edu.cn,
dongzhenwhu@whu.edu.cn, bshyang@whu.edu.cn

Abstract

3D open-vocabulary scene understanding, which accurately perceives complex semantic properties of objects in space, has gained significant attention in recent years. In this paper, we propose GAGS, a framework that distills 2D CLIP features into 3D Gaussian splatting, enabling open-vocabulary queries for renderings on arbitrary viewpoints. The main challenge of distilling 2D features for 3D fields lies in the multiview inconsistency of extracted 2D features, which provides unstable supervision for the 3D feature field. GAGS addresses this challenge with two novel strategies. First, GAGS associates the prompt point density of SAM with the camera distances to scene objects, which significantly improves the multiview consistency of segmentation results. Second, GAGS further decodes a granularity factor to guide the distillation process and this granularity factor can be learned in a unsupervised manner to only select the multiview consistent 2D features in the distillation process. Experimental results on two datasets show that GAGS improves visual grounding accuracy by an average of 10.9% and semantic segmentation accuracy by an average of 7.0%, with an inference speed 2× faster than baseline methods.

Code — https://github.com/WHU-USI3DV/GAGS

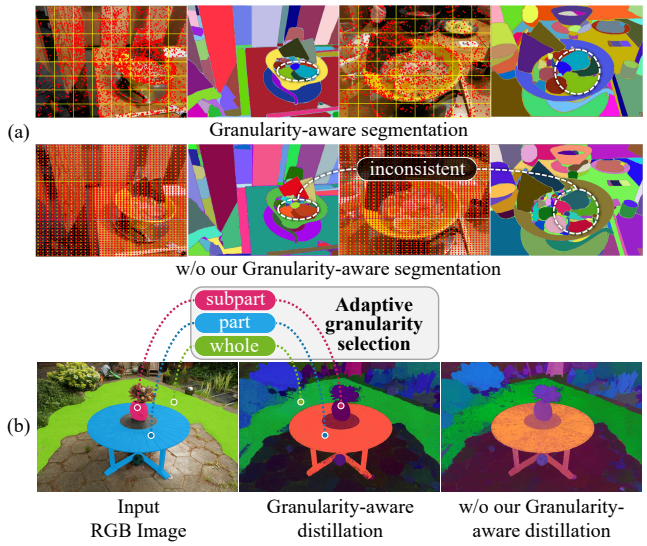

Figure 1: Our granularity-select feature learning strategy leverages the inherent consistency of 3D Gaussian splatting to perform granularity-aware feature distillation, enhancing the stability and accuracy of learned object features.

1 Introduction

3D scene understanding is a fundamental task in computer vision and a critical challenge in fields like robotics (Huang et al. 2023a; Shen et al. 2023; Wang et al. 2024) and autonomous driving (Jatavallabhula et al. 2023; Zheng et al. 2024; Cao et al. 2025). Recent advances in artificial intelligence and deep learning have driven research on open-vocabulary scene understanding, enabling users to query scene models using natural language (Peng et al. 2023; Ding et al. 2023; Wu et al. 2024a). This enhances the efficiency and intuitiveness of human-computer interaction, fostering closer integration between intelligent systems and human cognitive processes.

Due to the lack of large-scale, diverse 3D scene datasets with language annotations, current efforts focus on extending the knowledge of 2D vision-language models to 3D scenes. Early methods, such as OpenScene (Peng et al. 2023), compute pixel-level embeddings using pre-trained segmentation models and project them onto 3D point clouds. Recent approaches like LERF (Kerr et al. 2023) employ NeRF (Mildenhall et al. 2021) to represent 3D scenes, directly integrating CLIP (Radford et al. 2021) features into the scene modeling. Another recent work LangSplat (Qin et al. 2024) extends the open-vocabulary CLIP feature learning to the 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023), which *associates a learnable feature to each Gaussian kernel, supervises the rendered feature maps to be the same as the CLIP features extracted on each image.*

A noticeable challenge of learning such open-vocabulary features for 3D fields lies in the multiview inconsistency of the extracted 2D features. Brute-force training a 3D feature field from inconsistent multiview 2D features leads to blurred and degenerated 3D features. LangSplat (Qin et al. 2024) proposes to extract multi-view CLIP features by first segmenting images with SAM and argues that different

*These authors contributed equally.
†Corresponding author.

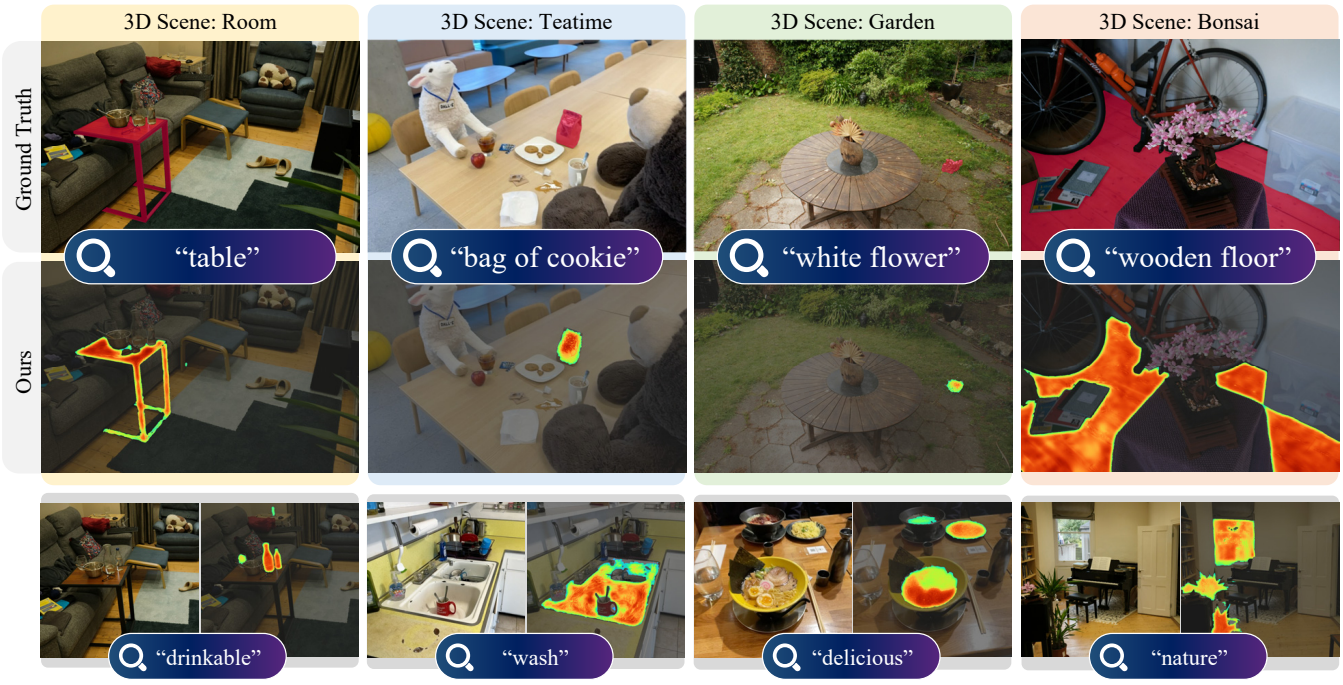

Figure 2: Given multiview images of a scene, GAGS learns a 3D Gaussian field associated with semantic features, which enables accurate open-vocabulary 3D visual grounding in the scene.

SAM segmentation granularities (i.e., whole, part, subpart) offer complementary multi-view consistency for objects at different scales. It thus extracts CLIP features from whole, part, and subpart segments and distills them into three separate feature fields to avoid missing consistent object features. However, this significantly increases training cost and query overhead, as each query must interact with all three fields. FastLGS (Ji et al. 2024) speeds up queries by only distilling multi-view feature indexes, yet still needs to store original multi-view multi-ganularity CLIP features for retrieval.

To reduce the storage and query overhead, N2F2 (Bhalgat et al. 2024) propose merges multi-granularity features by selecting the most activated ones based on predefined descriptions, while the simple descriptions cannot guarantee multi-view feature consistency. Other approaches (Ye et al. 2023; Wu et al. 2024b; Cen et al. 2025a) first segment objects in the Gaussian field, then detect target objects using grounding model outputs or 2D CLIP features. However, due to segmentation inaccuracies and conflicts among multi-view features, the final results still exhibit blurriness or errors.

In this work, we introduce Granularity-Aware 3D Feature Learning for Gaussian Splatting (GAGS), a framework to learn a *lightweight, unified*, and *highly consistent* feature field from multi-view, multi-granularity feature maps, enabling open-vocabulary queries as shown in Figure 2.

Our first intuition is that, for each granularity, consistent multi-view SAM segmentation serves as the foundation. SAM relies on the prompt points to determine the segmentation granularity. However, we find that naïve SAM segmentation with uniform or random prompt points leads to multi-view inconsistencies. As shown Figure 1(a), we improve the consistency of SAM segmentation results among multiview images by introducing a granularity-aware prompting strategy. We associate the prompt point density with the camera distance to the target object by utilizing the pre-trained 3D GS. Distant views receive dense prompts while nearby views use sparse prompts. We found that this adaptive prompting effectively results in consistent segmentation across multi-view images and further improves the CLIP feature consistency.

Then, on each image, we can extract CLIP features of three granularities, i.e. sub-parts, parts, and objects. Given that different segmentation granularities are suited to objects of different scales (Qin et al. 2024), the feature distillation process should allow each object to adaptively select an appropriate SAM granularity, i.e, *its SAM segmentation and corresponding CLIP feature are best consistent under this granularity.* To this end, we incorporate a granularity decoder on the learned Guassian Feature to decode a granularity factor as shown in Figure 1(b). We use this factor in supervision to help the distillation only learn the best granularity, i.e., with multiview consistent CLIP features, for various objects while neglecting inconsistent granularities. Note that the granularity scale itself is automatically learned without additional annotation or supervision.

We conduct experiments on an augmented LERF (Kerr et al. 2023) dataset and a self-annotated Mip-NeRF360 (Barron et al. 2022) dataset for evaluation. Experimental results demonstrate that our method outperforms baseline methods in open-vocabulary localization and semantic segmentation across all object granularities with a 2× improvement in query speed.

2 Related Work

2.1 Point-based Radiance Field

Radiance fields have long been used for 3D scene representation (Gortler et al. 1996; Levoy and Hanrahan 1996). Neural radiance fields (NeRF)(Mildenhall et al. 2021) significantly improved rendering quality by learning full scene representations via deep networks, enabling high-fidelity novel view synthesis. However, NeRF suffers from slow training and inference due to ray sampling and large MLPs. Despite efforts to improve efficiency(Müller et al. 2022; Fridovich-Keil et al. 2022), a trade-off between speed and quality remains. More recently, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) proposes a point-based alternative, replacing volumetric rendering with fast α-blending over discrete 3D Gaussians, achieving real-time rendering and direct geometry access. Building on these advantages, we adopt 3DGS as our base framework and extend it for open-vocabulary 3D scene understanding.

2.2 SAM for 3D Scene Segmentation

As one of the most impressive foundational vision models, Meta's SAM (Kirillov et al. 2023) has demonstrated exceptional zero-shot 2D segmentation capabilities. SAM supports flexible prompts, allowing it to generate multi-level masks for target objects based on inputs such as points, bounding boxes, and masks. Additionally, SAM offers the capability to automatically segment the entire image into multi-granularity masks, including whole, part and sub-part. Numerous methods have since emerged to extend SAM's capabilities to 3D space. Anything-3D (Shen, Yang, and Wang 2023) elevates SAM's segmentation to 3D, while Part123 (Liu et al. 2024) uses SAM's segmentation masks to reconstruct 3D models with high-quality segmented parts. Feature 3DGS (Zhou et al. 2024) and Gaussian Grouping (Ye et al. 2023) integrate SAM's features and object segmentation results at pre-selected single-granularity into 3D Gaussians, achieving high-quality segmentation of novel views and 3D scenes. Semantic Gaussian (Guo et al. 2024) utilizes various prompts to obtain instance-level segmentations for feature extraction. Considering that SAM can produce multi-granularity masks focusing on objects at different scales, some methods (Kim et al. 2024; Liang et al. 2024; Zhan et al. 2025; Cen et al. 2025a) integrates these masks into scene representations jointly or separately. Compared to the above methods that directly use the vanilla SAM, we attempt to enhance the consistency and reliability of multi-view segments in both mask generation and integration stage.

2.3 Open-vocabulary Scene Understanding

Scene understanding is a basic task in the field of computer vision. With the rapid advancement of deep learning, numerous methods (Wu et al. 2015; Chen, Chang, and Nießner 2020; Wang et al. 2021) have made significant progress across various subtasks of scene understanding. However, the limited availability of 3D training data has long posed a challenge for achieving comprehensive 3D scene understanding. Some methods (Wu et al. 2024b; Ye et al. 2023) attempt to align the outcomes of 2D grounding models (Ren et al. 2024) with 3D category-agnostic grouping approaches, thereby indirectly achieving open-vocabulary understanding. Vision foundation models like CLIP (Radford et al. 2021) have opened new avenues for open-vocabulary scene understanding. Due to the image-aligned nature of CLIP features, subsequent works such as CLIP2Scene (Chen et al. 2023) and Openscene (Peng et al. 2023) directly leverage CLIP-based 2D scene understanding models (Dong et al. 2023; Li et al. 2022) to obtain dense features with CLIP semantics. Approaches like (Zhang, Dong, and Ma 2023; Liu et al. 2023a; Kerr et al. 2023; Zuo et al. 2024; Shi et al. 2024) generate dense semantic features by performing multi-level image cropping and feature fusion, while often incorporating pixel-aligned feature supervision, such as DINO (Caron et al. 2021), to address the blurriness in semantic boundary. Some other researches (Kobayashi, Matsumoto, and Sitzmann 2022; Hong et al. 2023; Huang et al. 2023b; Qin et al. 2024; Peng et al. 2024; Cheng et al. 2024; Cen et al. 2025b) utilize pre-trained image segmentation model (Li et al. 2022; Cheng et al. 2022; Kirillov et al. 2023) to obtain semantically meaningful object-level patches, enabling more accurate scene understanding.

3 Method

Given multi-view images with camera poses, our goal is to learn a 3D feature field represented by 3D Gaussians with attached semantic features, enabling text-driven downstream tasks such as object localization and semantic segmentation.

An overview of our method is provided in Figure 3. We first train an RGB Gaussian field from the input images and attach a trainable low-dimensional semantic feature to each 3D Gaussian. We aim to enforce the consistency between multi-view rendering features and multi-view semantic features extracted by CLIP. To obtain pixel-level CLIP features for multi-view images, following LangSplat (Qin et al. 2024), we first segment each image with SAM and then extract CLIP features for each segmented region. The accuracy of the Gaussian feature field is highly sensitive to inconsistencies in multi-view CLIP features. To address this, we first propose GaS (Section 3.1), which corrects SAM prompt points to ensure consistent multi-view segmentation. SAM can provide three segmentation results with different granularities. To obtain a unified feature field, we need to select one granularity for various objects to extract and distill CLIP features. Therefore, we further propose GaD (Section 3.2), which adaptively selects SAM mask granularity and CLIP features with strong multi-view consistency during Gaussian feature distillation.

3.1 Granularity-aware Segmentation

The extracted CLIP features are strongly related to the region size of the segmentation results produced by SAM, different segmentation granularity leads to totally different CLIP features. However, since the SAM is applied to every input image separately and the same object may show different sizes on different viewpoints, the segmentation results of SAM are already multiview inconsistent, resulting in inconsistent CLIP features. For example, the ramen bowl in

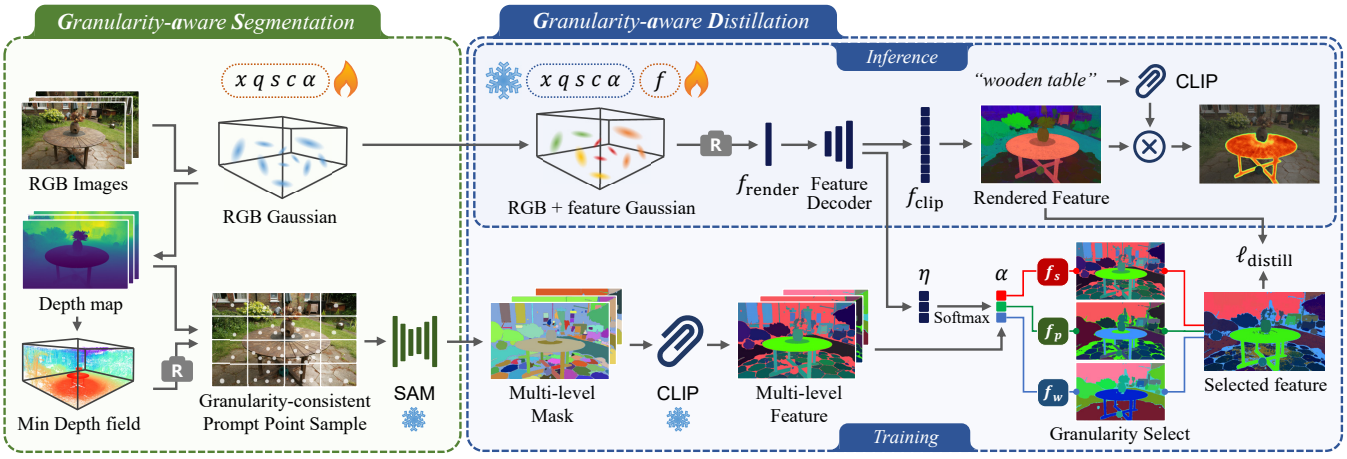

Figure 3: GAGS pipeline. *Left*: GAGS utilizes the scene's geometric representation to guide SAM and CLIP in producing multi-view consistent semantic features. *Right*: GAGS introduces a self-supervised granularity optimization strategy during feature distillation stage, which adaptively selects an appropriate feature granularity for each object in the scene.

Figure 1 corresponds to different segmentation granularities across various viewpoints. We propose a granularity-aware segmentation to improve the multiview consistency of segmentation results.

Observations The segmentation granularity of SAM is controlled by the prompt point density. In the ramen bowl region of Figure 1, dense prompt points usually lead to small segmentation regions while sparse prompt points result in large regions.

Another observation is that the apparent size of the same object in the image changes with the camera distance: it appears smaller when farther away and larger when closer. Consequently, the optimal prompt point density becomes depth-dependent—denser sampling is needed when the object appears smaller in distant views, while sparser sampling suffices when the same object appears larger in closer views.

Granularity-aware prompt point density We compute the prompt point density from the depth maps as follows. Given an input image, we begin by dividing it into a grid of patches and our target is to determine how many sample points should lie in each patch. Since the rendered depth determines the density, we first render a depth map D from the 3D Gaussian field.

Then, we back-project each depth map pixel to the 3D Gaussians and find the minimum depth value of each Gaussian after occlusion determination, denoted by MD, which maps a pixel to the minimum visible depth of its corresponding 3D position across all views. Then, the number of prompt points n_P for a specific patch P is computed by

$$n_P = \frac{1}{|P|} \sum_{p \in P} \frac{D^2(p)}{MD^2(p)} \cdot n, \tag{1}$$

where $|P|$ is the pixel number in the patch, $p \in P$ is a pixel, $MD^2(p)$ is the squared minimum depth value of pixel p while $D^2(p)$ is the squared depth value of pixel p, and n is a predefined prompt point number.

Explanation of Equation 1 The ratio $\frac{D^2(p)}{MD^2(p)}$ controls the prompt point density on a specific view. When this view is the one with the smallest viewing distance, the ratio is 1.0, which means we sample n points on the nearest view. While, for a far-away viewpoint, we use a larger prompt point density because the object would be smaller on these views, which can be seen in Figure 1. In this way, Equation 1 enables granularity-invariant prompting for the SAM model when segmenting the same object from different viewpoints, which increases the multi-view consistency of the segmentation results.

Moreover, after determining the number of prompt points per patch, the density distribution of visible gaussians is utilized to guide local prompt point sampling.

3.2 Granularity-aware Distillation

The above granularity-aware segmentation improves the consistency in SAM segmentations within each granularity. Then, we further propose a novel granularity-aware distillation to autonomously select the granularity best suiting each object, i.e., with the best multi-view CLIP consistency under this granularity. We represent three levels of SAM segmentation results by m_s, m_p and m_w, corresponding to "subpart", "part", and "whole". We then extract CLIP feature maps f_s, f_p and f_w on all three levels and design a strategy to let the distillation automatically select the most multiview-consistent features.

Decoding granularity factor Given the 3D Gaussian field associated with trainable feature vectors, we apply the splatting technique to render a feature map f_{render} for a training viewpoint. We adopt a feature decoder to decode a granularity factor and a predicted CLIP feature from the rendered feature vectors by

$$f_{\text{clip}}, \eta = \mathcal{D}(f_{\text{render}}), \tag{2}$$

where f_{clip} is the predicted CLIP feature, $\eta \in \mathbb{R}^3$ is the predicted granularity factor, and $\mathcal{D}$ means the MLP-based de-

coder shared among all viewpoints.

Weighted distillation loss We convert the granularity factor into weight by applying the softmax operator between three scores $\alpha = \text{softmax}(\eta)$. Then, we will train with a weighted L2 loss for distillation

$$\ell_{\text{distill}} = \sum_{n \in \{s,p,w\}} \alpha_n \|f_{\text{clip}} - f_n\|_2^2, \tag{3}$$

where n means the name, s, p, w correspond to "subpart", "part", and "whole" respectively, and $\alpha_n \in [0, 1]$ means the weight for the current level. We do not directly supervise α_n but let the optimization process automatically select the best weight to learn the granularity of the feature.

Entropy regularization To further encourage the weight to converge to a single scale instead of learning an averaged feature, we adopt an entropy regularization term

$$\ell_{\text{entropy}} = -\sum_{n \in \{s,p,w\}} \alpha_n \log \alpha_n, \tag{4}$$

which encourages the weight to be either 0 or 1. By adopting the granularity-aware distillation, GAGS only learns one 3D feature field instead of 3 feature fields in LangSplat (Qin et al. 2024), resulting in a more compact representation. The predicted granularity factor enables adaptively adjusting the granularity for different views and selecting the most multiview-consistent one.

Region-aware weighted distillation In our experiments, we observed that due to perspective variations and intrinsic object scale differences, different objects occupy varying proportions within the same image. During loss calculation, objects with a larger region in the image contribute more significantly to the loss, thereby dominating the optimization direction of semantic features. This imbalance may hinder feature learning for smaller objects. To address this issue, we propose a region-aware distillation loss, normalizing the loss by the region size of each object to ensure that all objects contribute equally to the total loss

$$\ell_{\text{r-distill}} = \beta_r \ell_{\text{distill}}, \beta_r = \frac{\sum_{i=1}^{n_r} S(R_i)}{n_r \cdot S(R)}, \tag{5}$$

where β_r denotes the region-aware factor, with n_r as the number of objects in the image and $S(R)$ as the area of each region. However, region-aware strategy is non-trivial. The core challenge here is constructing a mask that can simultaneously segment both large and small objects for balanced supervision, where SAM masks at any single granularity fails to achieve this as the whole scale mask m_w focuses on large objects, while the part m_p and subpart m_s masks emphasize small objects. Therefore, we propose integrating SAM masks at multiple granularities to adaptively select the appropriate scale for objects of different sizes. Specifically, we construct m_f, where each pixel selects the SAM granularity with the highest weight, i.e., $m_f = m_{\underset{i \in \{s,p,w\}}{\arg\max}\,(\alpha_i)}$. We then cluster pixels in m_f to objects for computing $S(R)$.

Feature consistency loss Inspired by the process of contrastive learning, we further introduce a feature consistency loss to encourage the features inside the same segmentation region to be consistent with each other

$$\ell_{\text{cons}} = \sum_{i=1}^{n_r} \sum_{p \in R_i} \frac{(f_{clip}^p - \overline{f_{R_i}})^2}{S(R_i)} \tag{6}$$

where f_{clip}^p and $\overline{f_{R_i}}$ represent the f_{clip} of pixel p and the average f_{clip} of the i-th region. By pulling together features of the same object region across each viewpoint, we ultimately enhance the feature consistency for the 3D Gaussians belonging to the same object.

In summary, the optimization loss $\mathcal{L}$ is

$$\mathcal{L} = \ell_{\text{r-distill}} + \lambda_{\text{entropy}} \ell_{\text{entropy}} + \lambda_{\text{cons}} \ell_{\text{cons}} \tag{7}$$

4 Experiments

4.1 Experimental Protocol

Datasets We use an augmented LERF dataset and a self-annotated Mip-NeRF360 dataset for evaluation. The LERF (Kerr et al. 2023) dataset comprises 3700+ phone-captured images from 14 different scenes. LangSplat (Qin et al. 2024) provides text descriptions and corresponding multi-view segmentation masks within four scenes: "ramen", "waldo_kitchen", "teatime", and "figurines", aiming to evaluate text-based object localization and segmentation. To better assess each method's performance on different granularities, we report experimental results separately for objects categorized by SAM segmentations as subpart, part, and whole. The Mip-NeRF360 (Barron et al. 2022) dataset offers multi-view images of 9 indoor and outdoor scenes with complex foreground and detailed background objects. We annotated four of the most complex scenes: "room" , "counters", "garden" and "bonsai" using the same format as the LERF dataset to evaluate model performance in challenging real-world scenarios.

Baselines We adopt recent relevant works including GS-Grouping (Ye et al. 2023), LEGaussian (Shi et al. 2024), GOI (Qu et al. 2024), and LangSplat (Qin et al. 2024) as our baseline methods. We employ GSplat (Ye et al. 2024) to construct initial Gaussian fields and evaluate all methods in the same setting for fair comparisons.

Metrics We follow LERF (Kerr et al. 2023) and LangSplat (Qin et al. 2024) to evaluate text-based 3D localization and segmentation accuracy on multi-view images. For localization, we evaluate the mean accuracy (mAcc) of the predicted locations falling within the ground truth bounding boxes. For segmentation, we assess the mean Intersection over Union (mIoU) between the predicted and the ground truth masks. All metrics are additionally reported with statistics divided by sample level (subpart, part, whole).

Implementation Details We utilized the SAM ViT-H (Dosovitskiy 2020) and OpenCLIP ViT-B/16 (Cherti et al. 2023) for segmentation and feature extraction. All experiments are conducted on a single RTX-4090 GPU.

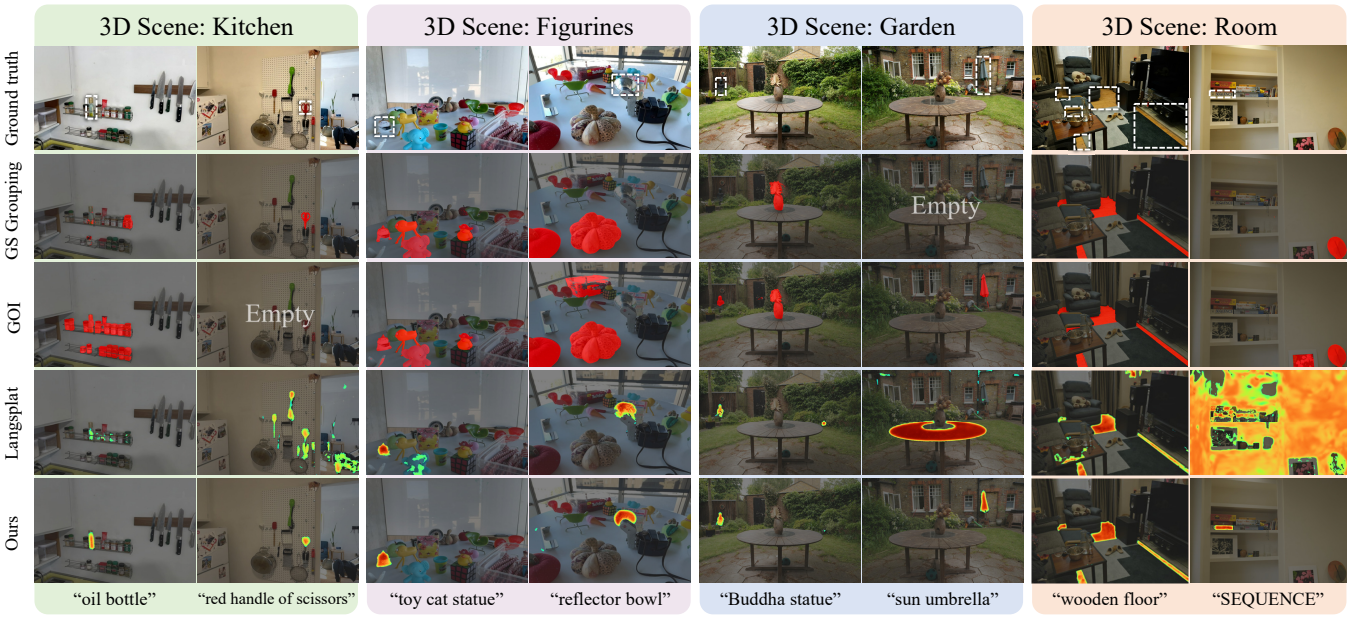

Figure 4: The relevance visualization results for open-vocabulary queries. Each row from top to bottom represents Ground Truth, Gaussian Grouping, GOI, Langsplat, and our method. Below each column is the corresponding input text description.

Method	**LERF**				**Mip-NeRF360**			
	Subpart	*Part*	*Whole*	*Overall*	*Subpart*	*Part*	*Whole*	*Overall*
GS-Grouping	11.67	27.97	58.77	36.08	17.5	57.47	80.44	57.09
LEGaussian	38.58	66.92	70.85	61.90	39.17	61.98	87.33	66.03
GOI	26.45	49.51	65.19	50.08	42.50	68.65	87.12	68.55
Langsplat	49.84	73.82	80.30	70.49	63.33	69.52	92.73	75.48
GAGS(**Ours**)	**51.39**	**87.93**	**88.17**	**79.14**	**78.97**	**95.83**	**100.00**	**88.67**

Table 1: Quantitative comparisons (mAcc, %↑) of 3D object location on the LERF and Mip-NeRF 360 dataset.

4.2 Comparison with Baseline Methods

Figure 4 shows segmentation results of baseline and our method. Table 1 and 2 report localization and segmentation accuracy on the LERF and Mip-NeRF360 datasets.

As shown in Figure 4, GS Grouping (Ye et al. 2023) and GOI (Qu et al. 2024) struggle to detect component-level and long-tail objects, which mainly due to ambiguities in feature field learning and the instability of 2D grounding model (Liu et al. 2023b) outputs. LangSplat (Qin et al. 2024) embeds multi-level CLIP features within 3 individual Gaussian fields, achieving better performance by avoiding erroneous feature average. However, querying across three Gaussian feature fields introduces more potential distractors, which can lead to incorrect retrievals.

In contrast, GAGS avoids the impact of both inconsistent multi-view features and features from inappropriate SAM segments by extracting and distilling consistent features into a unified Gaussian field. This results in improvements of 8.7% localization mAcc and 6.1% segmentation mIoU on the LERF dataset, and 13.2% mAcc and 7.9% mIoU on the Mip-NeRF360 dataset. Notably, GAGS shows larger gains at finer part levels on Mip-NeRF360 scenes, as small objects are more susceptible to severe occlusions in these complex scenes, making them particularly vulnerable to inconsistency or omission in multi-view SAM segmentation. Our GAGS, via GaS and GaD, can segments them more accurately while selecting effective and consistent CLIP features.

Runtime analysis We report the runtime of baselines and GAGS in Table 3. Although GOI (Qu et al. 2024) achieved the shortest training time, it requires fine-tuning a high-dimensional hyperplane during testing to segment the target, resulting in an inference time nearly 100 times that of GAGS. When comparing with LangSplat (Qin et al. 2024), GAGS achieves a similar training speed. However, LangSplat requires rendering multi-granularity feature maps to compare and segment the target, while GAGS adaptively fuses multi-granularity features, necessitating only a single feature map rendering and comparison. Thus, GAGS is two times faster than LangSplat in inference.

4.3 Ablation Studies

Table 4 presents ablation results on the Mip-NeRF360 dataset. Model 1, the baseline, using uniform prompt points for SAM and averaging CLIP features as $(f_s+f_p+f_w)/3$ for feature distillation. Incorporating Granularity-aware Distillation (GaD) in Model 2 boosts mAcc/mIoU by 9.4%/7.2%.

Method	LERF				Mip-NeRF360			
	Subpart	*Part*	*Whole*	*Overall*	*Subpart*	*Part*	*Whole*	*Overall*
GS-Grouping	7.94	25.54	52.66	31.67	15.77	47.13	70.51	49.22
LEGaussian	12.07	15.45	30.92	21.29	15.26	32.73	47.16	30.07
GOI	16.37	30.17	53.84	34.63	26.24	**57.99**	77.86	56.70
Langsplat	32.92	48.42	53.50	46.12	41.49	51.07	73.91	57.19
GAGS(**Ours**)	**35.89**	**55.09**	**59.52**	**52.21**	**56.07**	57.63	**79.6**	**65.09**

Table 2: Quantitative comparisons (mIoU, %↑) of 3D semantic segmentation on the LERF and Mip-NeRF 360 dataset.

Method	Training (min)				Inference (s)				
	Prep.	GS	Lang.	Total	IE	Rend.	Pred.	OSH	Total
GS-Grouping	20	10	40	70	66	~1	~1	-	68
GOI	10	10	10	**30**	-	~1	~1	2342	2344
Langsplat	50	10	30	90	-	15	34	-	49
GAGS(**Ours**)	25	10	50	85	-	13	11	-	**24**

Table 3: Time evaluation on the scene "ramen" of the LERF dataset. During training, "Prep." refers to the preprocessing stage, while "GS" and "Lang." represent the training stage of RGB and feature Gaussian fields, respectively. During inference, "IE" and "OSH" denotes obtaining the target's Identity Encoding and the optimization process of the Optimizable Semantic-space Hyperplane, while "Rend." and "Pred." refer to rendering features and predicting target mask.

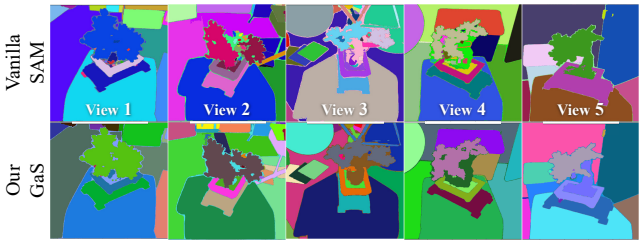

Figure 5: Visualization of multi-view segmentation results.

Figure 6: Visualization of learned granularity weights α.

ID	Setting GaS	Setting Distill.	Performance mIoU(%)	Performance mAcc(%)
1		Avg.	53.72	76.84
2		GaD	60.99	86.27
3	✓	**GaD**	**65.09**	**88.67**
4	✓	f_s	47.78	77.32
5	✓	f_p	61.68	85.74
6	✓	f_w	58.63	78.51
7	✓	Avg.	59.45	84.88

Table 4: Ablation Studies on the Mip-NeRF360 dataset. "Distill." denotes the feature distillation strategy. In addition to distilling the averaged features across all granularities, we also conducted experiments for each individual granularity feature. All other settings are kept the same.

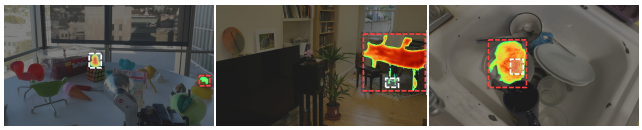

Figure 7: Visualization of failure cases. White boxes: ground truth; Red boxes: erroneous activation regions.

Model 3 with Granularity-aware Segmentation (GaS) yielding further gains of 2.4%/4.1% mAcc/mIoU. As shown in Figure 5, GaS achieves better multiview segmentation consistency by prompting SAM with fewer but proper seeds.

Models 4–6 distill CLIP features using single-granularity SAM segmentation. Among them, the part granularity performs best, while subpart performs worst due to the unreliable features and multi-view conflicts caused by oversegmentation of large objects. Model 7, which simply averages the CLIP features from all three SAM granularities, shows no improvement. In contrast, Model 3 with GaD can adaptively select reliable and multi-view consistent features. Finally, Model 3 achieves optimal performance, demonstrating the effectiveness of our proposed GaS and GaD modules.

5 Conclusion

Summary. We presented GAGS, a granularity-aware 3D Gaussian Splatting framework for efficient, accurate open-vocabulary scene understanding. By introducing a granularity-aware segmentation and feature distillation strategy, GAGS is able to learn clear and multi-view consistent semantics features. It avoids the high query cost of rendering multiple language fields and mitigates the impact of outliers. Experiments on several datasets demonstrate notable gains of GAGS in both accuracy and speed over prior methods.

Limitation. As Figure 7 shows, while self-supervised granularity optimization alleviates inference-time outlier issues, it may introduce information loss that limits component-level recognition in complex scenes. Future work will focus on mitigating this trade-off to further improve performance.

Acknowledgments

This work was supported by NSFC General Program (No.42571521), NSFC Program for Ph.D. (No.424B2012).

References

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5470–5479.

Bhalgat, Y.; Laina, I.; Henriques, J. F.; Zisserman, A.; and Vedaldi, A. 2024. N2F2: Hierarchical Scene Understanding with Nested Neural Feature Fields. *arXiv preprint arXiv:2403.10997*.

Cao, Z.; Mi, X.; Qiu, B.; Cao, Z.; Long, C.; Yan, X.; Zheng, C.; Dong, Z.; and Yang, B. 2025. Cross-modal semantic transfer for point cloud semantic segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221: 265–279.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Cen, J.; Fang, J.; Yang, C.; Xie, L.; Zhang, X.; Shen, W.; and Tian, Q. 2025a. Segment any 3d gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1971–1979.

Cen, J.; Zhou, X.; Fang, J.; Wen, C.; Xie, L.; Zhang, X.; Shen, W.; and Tian, Q. 2025b. Tackling View-Dependent Semantics in 3D Language Gaussian Splatting. *arXiv preprint arXiv:2505.24746*.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, 202–221. Springer.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cheng, J.; Zaech, J.-N.; Van Gool, L.; and Paudel, D. P. 2024. Occam's LGS: An Efficient Approach for Language Gaussian Splatting. *arXiv preprint arXiv:2412.01807*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Ding, R.; Yang, J.; Xue, C.; Zhang, W.; Bai, S.; and Qi, X. 2023. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7010–7019.

Dong, X.; Bao, J.; Zheng, Y.; Zhang, T.; Chen, D.; Yang, H.; Zeng, M.; Zhang, W.; Yuan, L.; Chen, D.; et al. 2023. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10995–11005.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; and Kanazawa, A. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5501–5510.

Gortler, S. J.; Grzeszczuk, R.; Szeliski, R.; and Cohen, M. F. 1996. The lumigraph. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, 43–54. New York, NY, USA: Association for Computing Machinery. ISBN 0897917464.

Guo, J.; Ma, X.; Fan, Y.; Liu, H.; and Li, Q. 2024. Semantic Gaussians: Open-Vocabulary Scene Understanding with 3D Gaussian Splatting. *arXiv preprint arXiv:2403.15624*.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36: 20482–20494.

Huang, C.; Mees, O.; Zeng, A.; and Burgard, W. 2023a. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 10608–10615. IEEE.

Huang, C.; Mees, O.; Zeng, A.; and Burgard, W. 2023b. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 10608–10615. IEEE.

Jatavallabhula, K. M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Maalouf, A.; Li, S.; Iyer, G.; Saryazdi, S.; Keetha, N.; et al. 2023. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*.

Ji, Y.; Zhu, H.; Tang, J.; Liu, W.; Zhang, Z.; Tan, X.; and Xie, Y. 2024. Fastlgs: Speeding up language embedded gaussians with feature grid mapping. *arXiv preprint arXiv:2406.01916*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kerr, J.; Kim, C. M.; Goldberg, K.; Kanazawa, A.; and Tancik, M. 2023. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19729–19739.

Kim, C. M.; Wu, M.; Kerr, J.; Goldberg, K.; Tancik, M.; and Kanazawa, A. 2024. Garfield: Group anything with radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21530–21539.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kobayashi, S.; Matsumoto, E.; and Sitzmann, V. 2022. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35: 23311–23330.

Levoy, M.; and Hanrahan, P. 1996. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, 31–42. New York, NY, USA: Association for Computing Machinery. ISBN 0897917464.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Liang, S.; Wang, S.; Li, K.; Niemeyer, M.; Gasperini, S.; Navab, N.; and Tombari, F. 2024. SuperGSeg: Open-Vocabulary 3D Segmentation with Structured Super-Gaussians. *arXiv preprint arXiv:2412.10231*.

Liu, A.; Lin, C.; Liu, Y.; Long, X.; Dou, Z.; Guo, H.-X.; Luo, P.; and Wang, W. 2024. Part123: Part-aware 3D Reconstruction from a Single-view Image. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Liu, K.; Zhan, F.; Zhang, J.; Xu, M.; Yu, Y.; El Saddik, A.; Theobalt, C.; Xing, E.; and Lu, S. 2023a. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36: 53433–53456.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

Peng, Q.; Planche, B.; Gao, Z.; Zheng, M.; Choudhuri, A.; Chen, T.; Chen, C.; and Wu, Z. 2024. 3d vision-language gaussian splatting. *arXiv preprint arXiv:2410.07577*.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 815–824.

Qin, M.; Li, W.; Zhou, J.; Wang, H.; and Pfister, H. 2024. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20051–20060.

Qu, Y.; Dai, S.; Li, X.; Lin, J.; Cao, L.; Zhang, S.; and Ji, R. 2024. GOI: Find 3D Gaussians of Interest with an Optimizable Open-vocabulary Semantic-space Hyperplane. *arXiv preprint arXiv:2405.17596*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

Shen, Q.; Yang, X.; and Wang, X. 2023. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*.

Shen, W.; Yang, G.; Yu, A.; Wong, J.; Kaelbling, L. P.; and Isola, P. 2023. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*.

Shi, J.-C.; Wang, M.; Duan, H.-B.; and Guan, S.-H. 2024. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5333–5343.

Wang, J.; Xu, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9401–9411.

Wang, J.; Zhang, Z.; Zhang, Q.; Li, J.; Sun, J.; Sun, M.; He, J.; and Xu, R. 2024. Query-based Semantic Gaussian Field for Scene Representation in Reinforcement Learning. *arXiv preprint arXiv:2406.02370*.

Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; et al. 2024a. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 5092–5113.

Wu, Y.; Meng, J.; Li, H.; Wu, C.; Shi, Y.; Cheng, X.; Zhao, C.; Feng, H.; Ding, E.; Wang, J.; et al. 2024b. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *Advances in Neural Information Processing Systems*, 37: 19114–19138.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Ye, M.; Danelljan, M.; Yu, F.; and Ke, L. 2023. Gaussian grouping: Segment and edit anything in 3d scenes. *arXiv preprint arXiv:2312.00732*.

Ye, V.; Li, R.; Kerr, J.; Turkulainen, M.; Yi, B.; Pan, Z.; Seiskari, O.; Ye, J.; Hu, J.; Tancik, M.; et al. 2024. gsplat: An Open-Source Library for Gaussian Splatting. *arXiv preprint arXiv:2409.06765*.

Zhan, C.; Zhang, Y.; Wang, G.; and Wang, H. 2025. Hi-LSplat: Hierarchical 3D Language Gaussian Splatting. *arXiv preprint arXiv:2506.06822*.

Zhang, J.; Dong, R.; and Ma, K. 2023. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2048–2059.

Zheng, W.; Song, R.; Guo, X.; and Chen, L. 2024. Genad: Generative end-to-end autonomous driving. *arXiv preprint arXiv:2402.11502*.

Zhou, S.; Chang, H.; Jiang, S.; Fan, Z.; Zhu, Z.; Xu, D.; Chari, P.; You, S.; Wang, Z.; and Kadambi, A. 2024. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21676–21685.

Zuo, X.; Samangouei, P.; Zhou, Y.; Di, Y.; and Li, M. 2024. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *International Journal of Computer Vision*, 1–17.