

Occam's LGS: A simple approach for Language Gaussian Splatting

Jiahuan Cheng^{1,2} Jan-Nico Zaech²

¹Johns Hopkins University
Baltimore, United States

Luc Van Gool² Danda Pani Paudel²

²INSAIT, Sofia University
Sofia, Bulgaria

Abstract

TL;DR: Gaussian Splatting is a widely adopted approach for 3D scene representation that offers efficient, high-quality 3D reconstruction and rendering. A major reason for the success of 3DGS is its simplicity of representing a scene with a set of Gaussians, which makes it easy to interpret and adapt. To enhance scene understanding beyond the visual representation, approaches have been developed that extend 3D Gaussian Splatting with semantic vision-language features, especially allowing for open-set tasks. In this setting, the language features of 3D Gaussian Splatting are often aggregated from multiple 2D views. Existing works address this aggregation problem using cumbersome techniques that lead to high computational cost and training time.

In this work, we show that the sophisticated techniques for language-grounded 3D Gaussian Splatting are simply unnecessary. Instead, we apply Occam's razor to the task at hand and perform weighted multi-view feature aggregation using the weights derived from the standard rendering process, followed by a simple heuristic-based noisy Gaussian filtration. Doing so offers us state-of-the-art results with a speed-up of two orders of magnitude. We showcase our results in two commonly used benchmark datasets: LERF and 3D-OVS. Our simple approach allows us to perform reasoning directly in the language features, without any compression whatsoever. Such modeling in turn offers easy scene manipulation, unlike the existing methods – which we illustrate using an application of object insertion in the scene. Furthermore, we provide a thorough discussion regarding the significance of our contributions within the context of the current literature. Project Page: <https://insait-institute.github.io/OccamLGS/>

Numquam ponenda est pluralitas sine necessitate.

William of Ockham

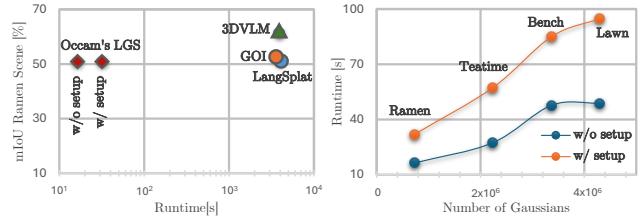


Figure 1. Occam's LGS performs training-free language feature aggregation in 3D by accurately modeling rendering of Gaussian Splatting, removing the need for expensive training in feature space. This improves runtime by two order of magnitudes, while achieving SOTA performance for open-set vision-language tasks. The straight-forward approach integrates language features seamlessly into the trained 3D Gaussian Splatting representation, allowing for downstream applications such as scene editing.

1. Introduction

3D semantic understanding is a fundamental capability in computer vision that enables machines to interpret and reason about the 3D world, with practical uses in robotics [27], autonomous navigation [33, 42], augmented reality, and interactive systems. While traditional 3D scene understanding [9] focuses on extracting geometric structure and fixed information from visual data, newer approaches integrate natural language to create richer and more meaningful scene interpretations. This emerging field, known as 3D language understanding [4, 7, 10, 13, 17, 23, 32, 41], aims to establish meaningful connections between linguistic expressions and 3D visual representations such as point clouds, meshes, and RGB-D images.

Recent breakthroughs in 3D scene representation [12, 19, 31, 39], particularly 3D Gaussian Splatting [12, 40], have revolutionized scene reconstruction and rendering by offering an efficient and high-quality approach to modeling 3D environments. While 3D Gaussian Splatting excels at geometric reconstruction and novel view synthesis, its potential for semantic understanding remains largely unexplored. Recent research has begun addressing this limitation by incorporating language features into the Gaus-

sian representations, enabling open-vocabulary scene understanding through natural language queries.

Vision-language models like CLIP [2] and LSeg [5] have demonstrated the power of joint visual-linguistic understanding by training on vast image-text pairs, while foundation models like Segment Anything (SAM) [14] and DINO-V2 [20] have shown the potential of large-scale visual learning. Although these models have changed 2D image understanding completely, extending their capabilities to 3D domains remains challenging due to limited 3D-language paired data. Recent approaches [1, 21, 24, 30, 36, 37] have attempted to bridge this gap by leveraging pre-trained 2D vision-language features for 3D understanding.

However, a fundamental challenge emerges when incorporating these 2D language features into 3D representations: 2D vision-language models typically produce high-dimensional feature vectors to capture rich semantic information and 3D scenes often contain more than one million Gaussians. Generating high-dimensional features for each 3D element using a training-based approach quickly becomes computationally infeasible.

Current approaches attempt to address this challenge through scene-specific feature autoencoders [24] or quantization to create a compact discrete feature space [30] that makes training feasible. However, these scene-specific compression approaches fundamentally limit scalability and generalization. While effective for individual environments, they create barriers for multi-scene applications by requiring separate training for each new scene, increasing both computational and storage overhead. Moreover, these scene-specific solutions make it difficult to transfer knowledge between environments and limit the potential for open-world understanding.

We take a fundamentally different approach by proposing a training-free method, that avoids repeated iteration over views, as well as loss computation in the high-dimensional feature space. Overall, our contribution can be summarized as follows:

- An training-free global optimization approach for Language 3D Gaussian Splatting, improving speed by two orders of magnitude.
- A reasoning by rendering approach to avoiding expensive feature rendering by accurately modeling the Language 3D Gaussian Splatting forward process.
- Compatibility to arbitrary language feature dimensionality, without feature compression.

2. Related works

3D Open Vocabulary Scene Understanding

Recent works [3, 6, 23, 34, 41] have explored leveraging 2D vision-language models such as CLIP [26], OpenSeg [5], and OVSeg [16], and foundation models such as SAM [14]

Method	Requires				Simple	Fast
	Feature Compression	Spatial Compression	Back Propagation	Scene Specific Model		
LERF [13]	✗	✗	✓	✓	✗	✗
GS-Grouping [37]	✓	✗	✓	✓	✗	✗
Feature-3DGS [43]	✓	✗	✓	✓	✗	✗
LEGaussians [30]	✓	✗	✓	✓	✗	✗
GOI [25]	✓	✗	✓	✓	✗	✗
LangSplat [24]	✓	✗	✓	✓	✗	✗
FMGS [44]	✗	✓	✓	✓	✗	✗
Occam’s LGS	✗	✗	✗	✗	✓	✓

Table 1. Our approach challenges the necessity of complex modules to lift 2D language features to Gaussian Splats, resulting in a simple and fast, yet effective method.

to ground 2D language-aligned features in 3D space. Initial NeRF-based approaches [11, 13, 15] demonstrated the feasibility of distilling language features into neural radiance fields, though at the cost of slow training and rendering. More recent Gaussian-based methods have improved efficiency while following a common framework: extracting language-aligned feature maps from 2D views and grounding them to 3D space through multi-view consistency. LangSplat [24] employs SAM to obtain masks at different granularity, then extracts CLIP image features for these regions. LEGaussians [30] combines DINO [2] and CLIP features to obtain hybrid dense language feature maps. FMGS [44] similarly leverages CLIP features with DINO regularization for more robust semantic representation. GOI [25] leverages APE [29] to extract pixel-aligned features with fine boundaries.

However, incorporating these pre-trained features into 3D space presents unique challenges, particularly in managing their high dimensionality. High-dimensional features not only increase memory requirements but also significantly extend training time during optimization. Different approaches have been proposed to address this challenge: LangSplat [24] employs a scene-specific autoencoder for feature compression, LEGaussians [30] uses quantization to discretize semantic features, and Feature-3DGS [43] employs a feature compression scheme with lightweight decoder. FMGS [44] introduces multi-resolution encoding for efficient semantic embedding. GOI [25] utilizes a trainable feature clustering codebook to condense high-dimensional semantic features.

In contrast to these methods, our approach supports arbitrary-dimensional feature distillation without compromising training speed. Furthermore, our method preserves

semantic meaning without requiring additional preprocessing steps or encoder-decoder structures for feature compression. This ensures generalization to the same extent as the underlying feature extractor and thus, enables a range of downstream applications requiring dynamic scenes or scene editing.

Uplifting 2D Features to 3D

The dominant paradigm in current methods [15, 21, 24, 30] involves augmenting each Gaussian with semantic features and optimizing them through iterative gradient descent. In addition, several approaches also explore the direct uplifting of 2D features to 3D. OpenGaussian [36] proposes an instance-level 3D-2D feature association method that links 3D points to 2D masks, leveraging the mIoU between rendered and ground truth masks. Semantic Gaussians [8] utilizes depth information to find corresponding 3D Gaussians along each pixel’s ray. PLGS [35] initializes semantic anchor points by backprojecting pixels into 3D space using camera parameters, while LUDVIG [18] aggregates multi-view features for each Gaussian by considering each pixel’s viewing direction.

While not focusing on semantic understanding, Flash-Splat [28] demonstrates that 2D mask rendering is essentially a linear function of Gaussian labels, enabling closed-form optimization through linear programming. These methods all rely on back-projection approaches, finding Gaussians along the pixel’s ray direction and using depth information for correspondence. However, such back-projection can be inaccurate due to occlusions and incomplete depth information.

In contrast, our method accurately models the Gaussian Splatting’s forward rendering to lift 2D language-aligned features to 3D, while avoiding the back-projection limitations. Table 1 summarizes the key differences between our method and existing approaches. Unlike previous approaches, our method is training-free and does not require costly explicit semantic feature rendering. While maintaining comparable performance with existing methods, it efficiently completes the lifting of 512-dimensional features to 3D space in less than one minute.

3. Method

The following section provides an overview of our method, which consists of 3 main steps, Reasoning by Rendering, Weighted Feature Aggregation, and Filtering, as depicted in Figure 2. An overview of the flow of our method is further given in Algorithm 1.

3D Gaussian Splatting represents a scene as a collection of anisotropic Gaussians, where each Gaussian is defined by its position $\mathbf{x} \in \mathbb{R}^3$, covariance $\Sigma \in \mathbb{R}^6$, color represented by either a 3D color representation $\mathbf{c} \in \mathbb{R}^3$, or a viewpoint dependent representation as 48 spherical harmonics (SH)

parameters, and opacity $o \in \mathbb{R}$

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x})^\top \Sigma^{-1}(\mathbf{x})}. \quad (1)$$

The primitive representation of a Gaussian is defined as $G = \{\mathbf{x}, \Sigma, \mathbf{c}, o\}$, and let $\mathcal{G} = \{G_i\}_{i=1}^N$ be the set of Gaussians in a scene. For rasterization, we first project each 3D Gaussian onto the image plane, resulting in a 2D Gaussian $G' = \{\mathbf{x}', \Sigma', \mathbf{c}, o\}$. Here, $\mathbf{x}' \in \mathbb{R}^2$ represents the projected 2D coordinates and $\Sigma' \in \mathbb{R}^{2 \times 2}$ is the projected covariance matrix.

To render the color at the pixel \mathbf{p} , we blend the contributions of all Gaussians along the ray originating from the camera center, say $\mathcal{G}^r = \{G_i\}_{i=1}^R$. To achieve efficient rendering in the Gaussian Splatting framework, this rendering process is implemented through a tile-based sorting strategy, that utilizes the approximation of Gaussians by a spatially truncated distribution. For this purpose, the image is partitioned into tiles, typically spanning 16x16 pixels. Each 2D Gaussian is assigned to the tiles intersecting with a bounding box placed at the 3σ boundary. Within each tile, Gaussians are sorted by depth from the camera center. These sorted Gaussians represent the collection of Gaussians that potentially contribute to any pixel within that tile, which are then rendered sequentially, starting closest to the camera. The tiling approach approximates the ray-Gaussian intersection test: if a Gaussian’s bounding box intersects a tile, it potentially contributes to any ray passing through that tile’s pixels.

For the i -th Gaussian in the sorted list, its contribution to a pixel \mathbf{p} is defined as $\alpha_i = o_i G_i(\mathbf{x}'_i)$, representing its influence. When projecting all relevant Gaussians, the color of the pixel \mathbf{p} is computed through a weighted blending process given by

$$\mathcal{C}(\mathbf{p}) = \psi_c(\mathbf{p}, \mathcal{G}) = \sum_{i=1}^R \alpha_i \mathbf{c}_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (2)$$

where, $\psi_c(., .)$ is the color rendering function. Specifically, the contribution of G_i to pixel \mathbf{p} is computed as

$$\alpha_i = \mathbf{o}_i \cdot e^{-\frac{1}{2}((\mathbf{p}-\mathbf{x}'_i)^\top \Sigma^{-1}(\mathbf{p}-\mathbf{x}'_i))}. \quad (3)$$

With the goal to lift multi-view 2D features into 3D space, we extend each Gaussian representation to include a high-dimensional semantic feature vector $\mathbf{f} \in \mathbb{R}^d$. Thus, each augmented 3D Gaussian G_i learns its own semantic feature vector \mathbf{f} . Similar to color blending, the semantic feature at pixel \mathbf{p} is computed through alpha-weighted blending

$$\mathcal{F}(\mathbf{p}) = \psi_f(\mathbf{p}, \mathcal{G}) = \sum_{i=1}^R \alpha_i \mathbf{f}_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (4)$$

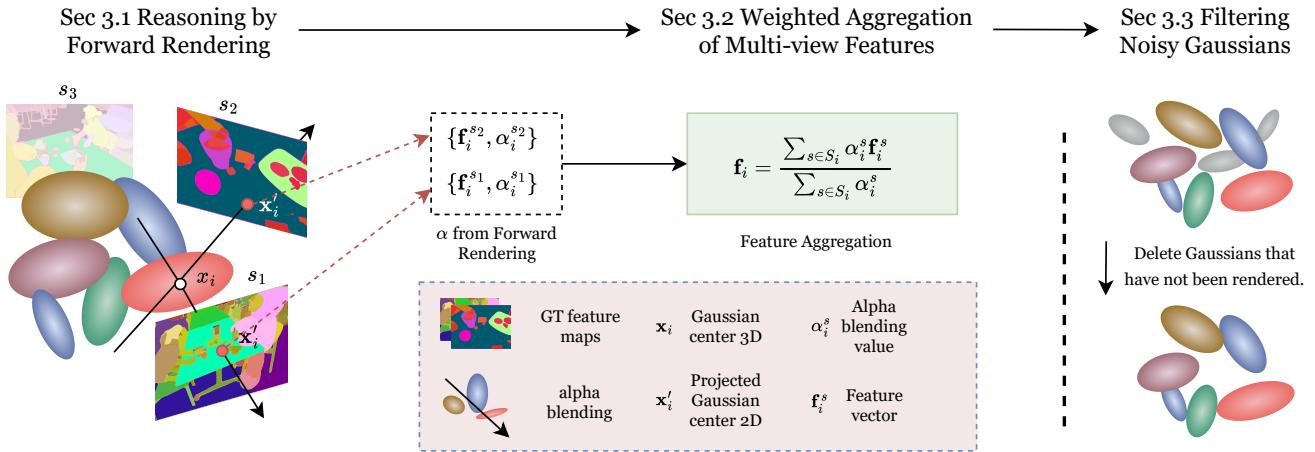


Figure 2. Overview of our method: Occam’s LGS consists of three main stages: (1) Forward rendering process of 3D Gaussian Splatting to obtain opacity α , projected positions x'_i for each Gaussian and their corresponding pixels p_i in 2D views (Sec. 3.1), (2) Weighted aggregation of multi-view semantic features using the alpha blending weights from the rendering process (Sec. 3.2), and (3) Filtering of noisy Gaussians that remain invisible throughout the rendering process (Sec. 3.3).

where ψ_f is the feature rendering function and α_i represents the same Gaussian contribution as defined in the color rasterization.

It is important to note that the rendering of semantics according to Equation (4) is expensive both in memory and computation when compared to the rasterization of color in Equation (2). This is due to the considerably higher dimensionality of the features $d \gg 3$, which prevents the GPU from fitting all required data in the shader L1 cache [24].

3.1. Reasoning by Forward Rendering

Conventional methods for uplifting 2D features typically rely on back-projection, where 2D pixel-space language features are projected back to 3D space through ray-tracing. Instead, we leverage the forward rendering design of Gaussian Splatting to achieve more precise feature lifting from 2D to 3D.

Given a trained visual Gaussian model, let S_i be the set of views where Gaussian G_i is visible (lies within the view frustum), and \mathbf{p}_i^s denote its projected pixel location in view s . During the forward rendering pass, we obtain the 2D features \mathbf{f}_i^s at each projected pixel location \mathbf{p}_i^s . The 3D semantic feature \mathbf{f}_i for Gaussian G_i can then be computed as:

$$\mathbf{f}_i = \mathcal{W}(\mathbf{f}_i^s | s \in S_i) \quad (5)$$

where \mathcal{W} is the weighted aggregation function as detailed in Section 3.2.

For an efficient implementation, the center locations of Gaussians are recorded during the tile-based rendering process. Specifically, for Gaussian G_i with projected 2D center

\mathbf{x}'_i , its projected pixel location \mathbf{p}_i is determined by

$$\mathbf{p}_i = \lfloor \mathbf{x}'_i \rfloor, \quad (6)$$

where $\lfloor \cdot \rfloor$ denotes the floor operation to obtain the pixel coordinates.

This forward feature collection approach offers several key advantages. First, the feature uplifting requires only a single forward pass through the views, making it computationally efficient. Second, view frustum culling and tile-based sorting naturally handle occluded Gaussians, ensuring we only process visible features. Finally, by explicitly recording projection locations during rendering, we establish direct and precise correspondence between 2D features and 3D Gaussians, eliminating the need for complex back-projection or iterative optimization methods.

3.2. Weighted Aggregation of Multi-View Features

Simple aggregation of multi-view features can introduce noise into semantic representations, particularly when some views have minimal Gaussian contributions. A Gaussian might be barely visible or heavily occluded in certain views, making their feature contributions less reliable. To address this, we consider the Gaussian’s contribution in each view during feature aggregation. For each Gaussian G_i , we leverage the contribution α_i^s at its directly projected pixel location \mathbf{p}_i^s in view s as the weighting factor. The weighted feature aggregation function can then be expressed as

$$\mathbf{f}_i = \mathcal{W}(\mathbf{f}_i^s, \alpha_i^s | s \in S_i) = \frac{\sum_{s \in S_i} \alpha_i^s \mathbf{f}_i^s}{\sum_{s \in S_i} \alpha_i^s}. \quad (7)$$

This weighted aggregation ensures that views, where the Gaussian has stronger visibility and influence, contribute more significantly to its final semantic feature, resulting in more robust and consistent 3D feature representations.

3.3. Filtering Noisy Gaussians

After the forward rendering pass, we identify and filter out Gaussians that have no contributions across all views. This filtering effectively reduces noise while reducing the number of Gaussians significantly. Through this process, we obtain a refined set of semantic Gaussians $\mathcal{G}^s = \{G_i^s\}_{i=1}^M$, where $M \ll N$, representing only the Gaussians that meaningfully contribute to the scene’s semantic representation.

Algorithm 1 Occam’s LGS

Require:

- 1: Training camera views $\mathcal{S} = \{s_1, \dots, s_n\}$
- 2: Set of trained 3D Gaussians $\mathcal{G} = \{G_1, \dots, G_m\}$
- 3: 2D language feature maps $\mathcal{F} = \{\mathbf{f}^{s_1}, \dots, \mathbf{f}^{s_n}\}$

Ensure:

- 4: **Initialize:**
 - 5: **for** each Gaussian $G_i \in \mathcal{G}$ **do**
 - 6: $\mathbf{f}_i \leftarrow 0$ {Accumulated features}
 - 7: $\alpha_i \leftarrow 0$ {Accumulated weights}
 - 8: **end for**
 - 9: **Feature Uplifting Phase:**
 - 10: **for** each camera view $s \in \mathcal{S}$ **do**
 - 11: **for** each Gaussian G_i in view frustum **do**
 - 12: $\alpha_i^s, \mathbf{p}_i^s \leftarrow \text{Rasterize}(G_i|s)$ {Project and rasterize}
 - 13: $\mathbf{f}_i^s \leftarrow$ 2D feature map for view s
 - 14: $\mathbf{f}_i \leftarrow \mathbf{f}_i + \alpha_i^s \mathbf{f}_i^s$ {Sampled feature}
 - 15: $\alpha_i \leftarrow \alpha_i + \alpha_i^s$ {Alpha blending value}
 - 16: **end for**
 - 17: **end for**
 - 18: **for** each Gaussian $G_i \in \mathcal{G}$ **do**
 - 19: $\mathbf{f}_i \leftarrow \mathbf{f}_i / \alpha_i$ {Feature Normalization}
 - 20: **end for**
 - 21: **return** \mathcal{G}
-

4. Experiments

We evaluate our method against other state-of-the-art 3D Language Gaussian Splatting methods on open vocabulary segmentation and localization tasks, as well as compare their computational efficiency metrics.

4.1. Dataset

To assess the effectiveness of our method, we evaluate it on two widely used datasets: LERF [13] and 3D-OVS [17]. The LERF dataset features everyday indoor scenes with diverse objects and we perform quantitative evaluations following the on the commonly used scenes Ramen, Figurines,

Teatime, and Waldo Kitchen according to the LERF evaluation protocol [13]. For these scenes, we evaluate both mean Intersection over Union (mIoU) and localization accuracy. The 3D-OVS dataset contains long-tail objects in diverse backgrounds, and we evaluate the mIoU on five scenes: sofa, lawn, bed, room, and bench. For evaluation, we utilize the annotations provided by [24] for both datasets.

4.2. Implementation Details

Our implementation of Occam’s LGS is built on gsplat [38], a CUDA-accelerated Gaussian rasterization library that supports high-dimensional feature rendering. For each scene, we train a 3D Gaussian Splatt for 30,000 iterations using the common Gaussian Splatting [12] parameters, resulting in an average number of 2,500,000 Gaussians for RGB representation across scenes. To ensure a fair comparison, we follow the 2D language feature map generation method from [24]. For language feature extraction and segmentation, our pipeline integrates OpenCLIP ViT-B/16 [26] and SAM ViT-H [14]. To capture semantic information hierarchically, we generate 2D feature maps at three levels: whole, part, and subpart, where each level represents distinct semantic granularity. Following the feature extraction, we use our method to uplift the 2D language features to 3D scenes using training views only, and evaluate the results on separate testing views, withheld during training of the semantic representation. All experiments and runtime measurements are conducted on a single NVIDIA A6000 GPU.

For the LERF dataset, we follow the LERF [13] evaluation protocol, computing relevancy scores between query text and rendered 2D feature maps with a mask threshold of 0.5. For 3D-OVS, we first normalize the rendered feature maps to ensure comparable relevancy scores across classes, then apply dynamic thresholding detailed in the supplementary material. During thresholding, we reject spurious masks with areas smaller than 0.005% or larger than 0.90% of the image and set an object stability threshold of 0.4.

4.3. Open Vocabulary Semantic Segmentation

Given a 3D scene and a language query describing an object or region of interest, the task of open vocabulary semantic segmentation is to identify and generate masks for all relevant regions that match a query description. We evaluate our approach against existing methods on two datasets: LERF and 3D-OVS, with quantitative results shown in Tables 2 and 3 respectively and qualitative results shown in Figure 3 and 4. On the LERF dataset, under identical experimental settings as LangSplat, our approach demonstrates a substantial improvement of 9.9% mIoU. Notably, our method shows advantages, especially in complex scenes like Waldo Kitchen, where the camera captures an entire room rather than focusing on a specific area surrounded by camera views. Thus, these challenging scenarios come with

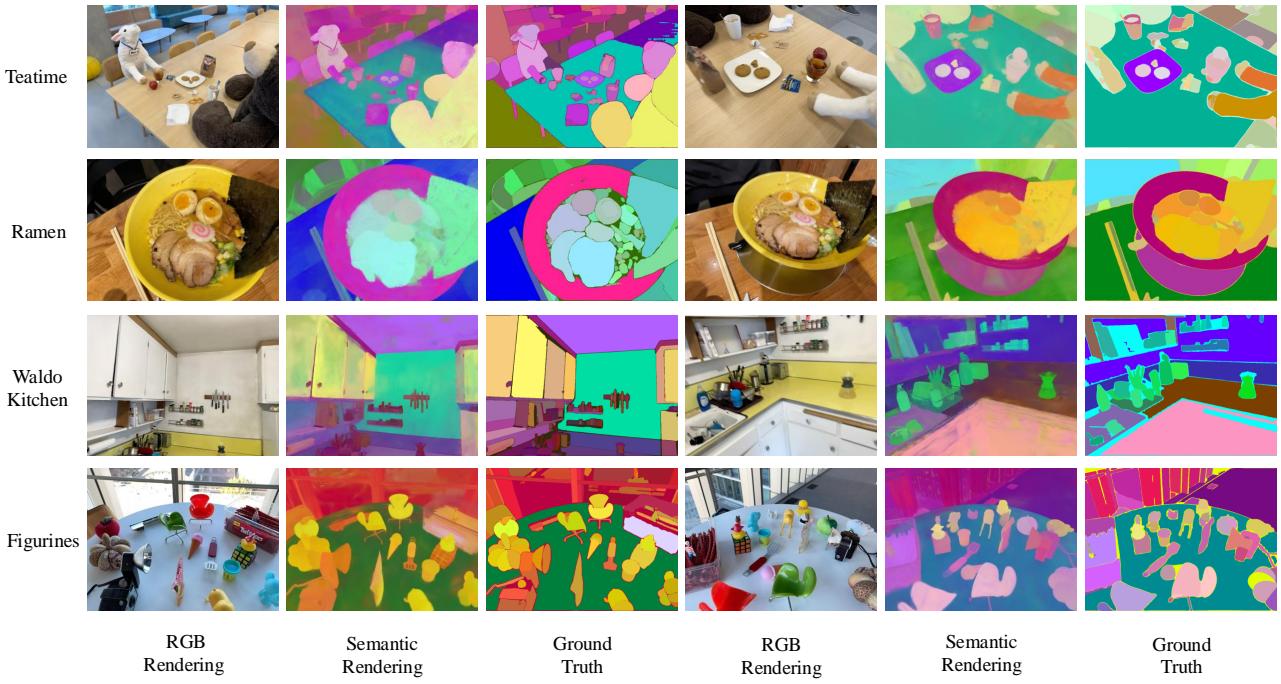


Figure 3. PCA visualization of semantic features on LERF Dataset. (a) Original RGB rendering. (b) PCA visualization of our rendered semantic feature maps. (c) Initial semantic feature maps extracted using CLIP from SAM segmentations. Note that colors are not aligned across different pairs of images due to independent PCA projections.

Method	Ramen	Figurines	Teatime	Waldo Kitchen	Overall
LSeg	7.0	7.6	21.7	29.9	16.6
LERF	28.2	38.6	45.0	37.9	37.4
GS-Grouping	45.5	60.9	40.0	38.7	46.3
Feature-3DGS	43.7	58.8	40.5	39.6	45.7
LEGaussians	46.0	60.3	40.8	39.4	46.9
GOI	52.6	63.7	44.5	41.4	50.6
LangSplat	<u>51.2</u>	44.7	<u>65.1</u>	<u>44.5</u>	<u>51.4</u>
Ours	51.0	58.6	70.2	65.3	61.3

Table 2. Comparison of average IoU on LERF Dataset.

sparse view coverage, compared to the size of the scene.

On 3D-OVS, our method achieves state-of-the-art results, with an improvement of 1.6% mIoU over the current SOTA LangSplat [24]. Performance gains of 3D-OVS are less pronounced compared to LERF [13], likely since this dataset contains simpler scenes, focused on a single region with good view coverage and fewer objects, where existing methods have already saturated in performance.

4.4. Localization

We further evaluate on the language-guided localization task, where given a language query, the goal is to local-

Method	Bed	Bench	Room	Sofa	Lawn	Overall
LSeg	56.0	6.0	19.2	4.5	17.5	20.6
LERF	73.5	53.2	46.6	27	73.7	54.8
GS-Grouping	83.0	91.5	85.9	87.3	90.6	87.7
Feature-3DGS	83.5	90.7	84.7	86.9	93.4	87.8
LEGaussians	84.9	91.1	86.0	87.8	92.5	88.5
GOI	89.4	92.8	91.3	85.6	94.1	90.6
LangSplat	<u>92.5</u>	<u>94.2</u>	<u>94.1*</u>	90.0	<u>96.1</u>	<u>93.4</u>
Ours	96.8	95.8	96.5*	<u>88.8</u>	97.0	95.0

Table 3. Comparison of average IoU on 3D-OVS Datasets. The utilized image resolution is 1440×1080 . Methods with * are evaluated on testing-data that has an annotation error in the 3D-OVS testset corrected.

ize the described object in 3D space. Following [13], we select the pixel with the highest relevancy score as the location of the object. Benchmarking against existing methods, our approach achieves comparable localization accuracy, as shown by Table 4.

4.5. Computational Efficiency

We evaluate the computational efficiency of our method against existing approaches in Figure 1. Similar to ours, LangSplat and GOI both build on pre-trained 3D Gaussian

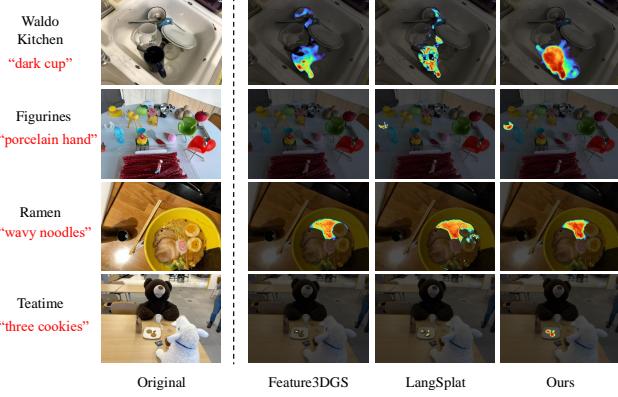


Figure 4. Qualitative comparison of relevancy score visualization. Trained with the same feature maps generated using CLIP and SAM.

Method	Ramen	Figurines	Teatime	Waldo Kitchen	Overall
LSeg	14.1	8.9	33.9	27.3	21.1
LERF	62.0	75.0	84.8	72.7	73.6
GS-Grouping	68.6	75.0	74.3	88.2	76.5
Feature-3DGS	69.8	77.2	73.4	87.6	77.0
LEGaussians	67.5	75.6	75.2	90.3	77.2
GOI	75.5	88.6	82.9	90.4	84.4
LangSplat	73.2	80.4	88.1	95.5	84.3
Ours	74.7	80.4	93.2	81.8	82.5

Table 4. Comparison of localization accuracy on LERF Datasets.

models, and we trained GOI using identical feature maps as LangSplat at whole object, part, and subpart granularities. For these existing methods, we report their semantic feature training times. LangSplat takes 67 minutes to train three levels of feature maps, while GOI takes 60 minutes of training time. With our approach, language features can be integrated in 16.35 seconds without setup (including checkpoint loading and saving), and 31.93 seconds with setup. In contrast, 3DVLM cannot utilize pre-trained Gaussians and thus, the reported time reflects their complete joint training of both reconstruction and semantics, which takes 65 minutes [22]. Thus, our method, which allows to use pre-trained visual Gaussians achieves semantic feature integration two orders of magnitude faster than other methods, while delivering superior mIoU performance across test datasets.

To analyze the scalability of our method, we evaluate the relationship between training time and the number of Gaussians present in a scene. For this purpose, we test our approach on four scenes, each containing a different number of Gaussians due to varying scene complexity and resolutions. The results depicted in Figure 1 demonstrate that our method scales well and maintains high computational efficiency even with large scenes containing up to 4 Million

Method	Ramen	Figurines	Teatime	Waldo Kitchen	Overall
Ours	51.0	58.6	70.2	65.3	61.3
(w/o wtd. agg)	45.3	56.7	70.0	64.8	59.9
(w/o filter)	47.9	57.1	67.7	62.3	58.8
Baseline	44.1	54.9	69.9	59.6	57.1

Table 5. Ablation study on the impact of filtering and weighted aggregation on feature learning. Evaluate on LERF dataset and use average IoU as the metric for comparison.



Figure 5. Qualitative results from our ablation study. Top: feature map visualization using our full method with filtering and weighted aggregation. Bottom: baseline results without either component.

Gaussians, indicating good scalability in practice.

4.6. Ablation Study

We validate the key components of our method with an ablation studies depicted in Table 5 and Figure 5. First, we ablate our fundamental choice to model the forward rendering process with the weighted feature aggregation strategy. Removing it, while keeping the Gaussian filtering process (w/o wtd. agg), shows that weighted aggregation improves the fusion of multi-view features due to the accurate modeling. Second, we analyze the impact of filtering redundant Gaussians that are not activated during rendering. Removing filtering (w/o filter) decreases performance, which supports the choice to filter inactive Gaussians, besides the observation that it enhances the smoothness and improves computational efficiency. Finally, we compare to a baseline not using weighting and filtering and thus, perform basic back-projection. This reduces performance by 4.2%, and altogether demonstrates that both components contribute meaningfully to our method’s performance.

4.7. Application: Object Insertion

A major advantage of the proposed method is the deep integration of dense 3D language features in 3D Gaussian Splatting and the inherent simplicity of its feature representation. This allows us to directly profit from any advances in 2D language feature representations as well as from the open-set generalization capability of these methods. Therefore, our approach has the potential to enable a range of down-

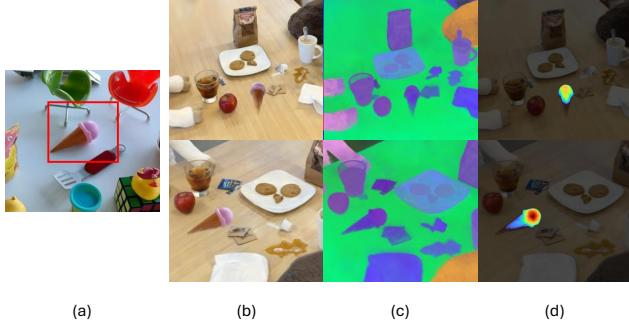


Figure 6. Visualization of Object Insertion: (a) Object Selection: A specific object is selected in the Figurines scene (b) Object Insertion: The selected object is transferred and integrated into the Teatime scene (c) Feature Analysis: PCA visualization of the feature map to show the object’s integration in the semantic space (d) Query Visualization: Heatmap showing the response to the query “pink icecream”.

stream 3D tasks, especially related to scene editing and interaction.

We demonstrate this capability on object insertion, where an object is taken from an existing 3D Gaussian Splatt and integrated in another scene. While the current state-of-the-art LangSplat would require finetuning of the scene-specific language autoencoder, as well as the retraining of the whole feature embedding, our representation can be augmented with the new object similarly to any other Gaussian splat. As illustrated in Figure 6, we select and extract an ice cream cone (represented by its Gaussians) from the Figurines scene of LERF [13]. By simply copying the object’s Gaussians together with their parameters and semantic features, the new object seamlessly integrates into the Teatime scene while preserving its semantic features.

To verify the successful integration, we query the modified scene with “pink icecream”. The resulting relevancy heatmap in Figure 6(d) confirms that our method maintains the semantic understanding of the inserted object in its new context. This demonstrates that, unlike existing methods that require scene-specific dimensional reduction techniques, our approach enables direct manipulation of semantic features, making it more flexible for practical scene editing operations while preserving semantic querying capabilities.

5. Discussion and Conclusion

In this work, we presented an optimization-free approach for augmenting Gaussian Splatting with language features. The proposed method shows state-of-the-art results with a single iteration over all views, and avoids rendering and loss computation in the feature space.

Beyond the immediate performance gains, our approach opens up new possibilities for how we approach 3D scene

understanding and manipulation. The ability to reason directly in language feature space, without compression, enables practical applications like object insertion and scene registration that were previously computationally prohibitive. Overall, we demonstrated that:

- High-dimensional language features can directly be used in Gaussian splatting by utilizing a training free optimization approach.
- Accurate modeling of the rendering process enables the backproject of highly relevant features from 2D to 3D.
- Directly projecting high-dimensional features allows us to use scene editing approaches previously limited to 3D Gaussian Splatting.

Based on the experiments and results presented in this work, we can answer the following questions about our work and its relation to existing works using feature compression:

Is feature compression or spatial mapping necessary for Language Gaussian Splatting? No. Avoiding the explicit rendering of language features and loss computation in the 2D space together with a training-free approach makes the direct use of high-dimensional language features feasible. Avoiding compressions algorithms significantly reduces training time and helps generalization of Language Gaussian Splatting.

Should we stop using low-dimensional Language Gaussian Splatting? No. Low dimensional features are justified in certain scenarios with consideration of their constraints. Though our method enables the direct use of uncompressed features, low-dimensional features can reduce the memory footprint of any 3D representation that relies on language features and speed up rendering of 3D featuremaps. However, it needs to be considered that using low-dimensional embeddings contradicts generalization across scenes, if compression approaches are trained scene-specifically.

Are there applications where low-dimensional language features are beneficial? Yes. For scenarios where the domain of scenes is constrained, no scene editing is required, and storage and rendering cost exceeds the cost of training a feature compressor, low dimensional features are advantageous. Thus, the optimal feature resolution is strongly application dependent and scales especially with the domain of potential scenarios.

Overall, it is important to acknowledge that our approach is complementary to existing works in Language Gaussian Splatting and enables a new direction to train the 3D language representation. Due to this complementary nature, Occam’s LGS can be combined with existing feature compression approaches, with the potential to speed up rendering and reduce storage cost. The right combination of approaches depends on the application and expected usecase of the 3D representation, which is a direction we hope future work will explore.

References

- [1] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. N2f2: Hierarchical scene understanding with nested neural feature fields. *arXiv preprint arXiv:2403.10997*, 2024. [1](#)
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. [2](#)
- [3] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023. [2](#)
- [4] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7010–7019, 2023. [1](#)
- [5] 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](#)
- [6] Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and PJ Narayanan. Interactive segmentation of radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4201–4211, 2023. [2](#)
- [7] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Concept-graphs: Open-vocabulary 3d scene graphs for perception and planning, 2023. [1](#)
- [8] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting, 2024. [3](#)
- [9] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112: 133–149, 2015. [1](#)
- [10] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. *arXiv preprint arXiv:2207.11514*, 2022. [1](#)
- [11] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19740–19750, 2023. [2](#)
- [12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. [1, 5](#)
- [13] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerp: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. [1, 2, 5, 6, 8](#)
- [14] 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. [2, 5](#)
- [15] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022. [2, 3](#)
- [16] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. [2](#)
- [17] 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. [1, 5](#)
- [18] Juliette Marrie, Romain Ménégaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes, 2024. [3](#)
- [19] 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. In *ECCV*, 2020. [1](#)
- [20] Maxime Oquab, Timothée Darct, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. [2](#)
- [21] Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and Ziyan Wu. 3d vision-language gaussian splatting. *arXiv preprint arXiv:2410.07577*, 2024. [2, 3](#)
- [22] Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and Ziyan Wu. 3d vision-language gaussian splatting, 2024. [7](#)
- [23] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies, 2023. [1, 2](#)
- [24] 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, 3, 4, 5, 6](#)

- [25] 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, 2024. [2](#)
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. [2, 5](#)
- [27] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans, 2020. [1](#)
- [28] QiuHong Shen, Xingyi Yang, and XinChao Wang. FlashSplat: 2d to 3d gaussian splatting segmentation solved optimally. In *European Conference of Computer Vision*, 2024. [3](#)
- [29] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13193–13203, 2024. [2](#)
- [30] 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, 3](#)
- [31] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5459–5469, 2022. [1](#)
- [32] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Open-mask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. [1](#)
- [33] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models, 2024. [1](#)
- [34] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3D distillation of self-supervised 2D image representations. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2022. [2](#)
- [35] Yu Wang, Xiaobao Wei, Ming Lu, and Guoliang Kang. Plgs: Robust panoptic lifting with 3d gaussian splatting, 2024. [3](#)
- [36] 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. *arXiv preprint arXiv:2406.02058*, 2024. [2, 3](#)
- [37] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179. Springer, 2025. [2](#)
- [38] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *arXiv preprint arXiv:2409.06765*, 2024. [5](#)
- [39] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinzhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxtels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2(3):6, 2021. [1](#)
- [40] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19447–19456, 2024. [1](#)
- [41] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2048–2059, 2023. [1, 2](#)
- [42] Hui Zhou, Xinge Zhu, Xiao Song, Yuexin Ma, Zhe Wang, Hongsheng Li, and Dahua Lin. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv preprint arXiv:2008.01550*, 2020. [1](#)
- [43] 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. [2](#)
- [44] Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *International Journal of Computer Vision*, pages 1–17, 2024. [2](#)

Occam's LGS: A simple approach for Language Gaussian Splatting

Supplementary Material

6. Evaluation

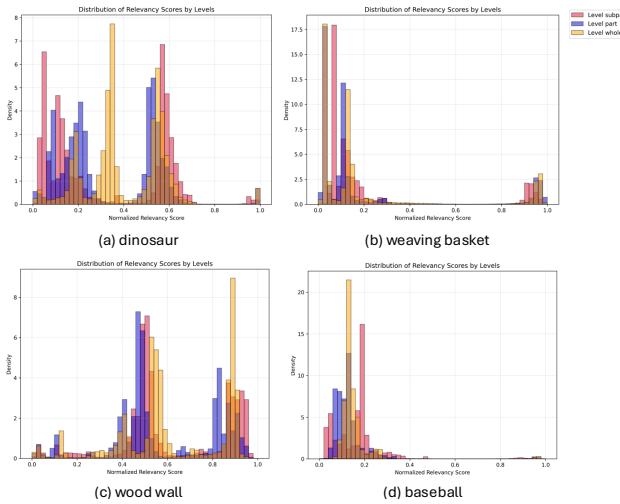


Figure 7. Distribution of relevancy scores across hierarchical levels (subpart, part, and whole) when applying different queries to room scene in the 3D-OVS dataset. Each color represents a distinct hierarchical level, illustrating the distribution of relevancy scores at different hierarchies.

We observe that different object queries yield varying ranges of relevancy scores, which makes it challenging to set a consistent threshold for object segmentation across query types, as shown in Figure 7. To address this limitation, we propose a dynamic thresholding approach. In other words, each query has a different threshold value. For each query, we evaluate different threshold values by incrementing with a step size of 0.01. At each threshold, we create a binary mask by selecting regions where the relevancy score exceeds the current threshold value, and compute the average relevancy score within this masked area. These average relevancy scores at different thresholds are plotted in Figure 9, shown by the increasing trend lines.

Our goal is to identify a "stable region" where variations in the threshold minimally affect both the mask size and average relevancy score. Specifically, we find the region where the absolute gradient of the average relevancy score (the rate of change of average relevancy score with respect to different thresholds) is below the chosen stability threshold, as shown in Figure 8. We then select the middle value from this stable range as our final threshold for each different query, as illustrated in Figure 9. The presence of a stable region indicates an object with consistent relevancy scores across threshold changes. In cases where multiple stable

regions exist, we select the rightmost one as it typically represents the highest confidence predictions. We apply this improved evaluation methodology to the 3D-OVS dataset

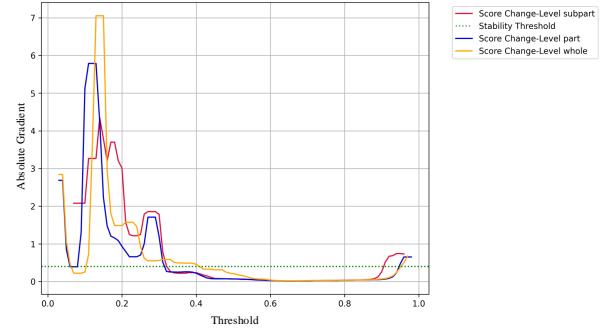


Figure 8. Visualization of the absolute gradient of average relevancy scores. The gradient measures the rate of change in average relevancy scores across different thresholds. The green dotted line indicates the stability threshold, with lower gradient values corresponding to more stable regions. This plot corresponds to the "weaving basket" query in room scene.

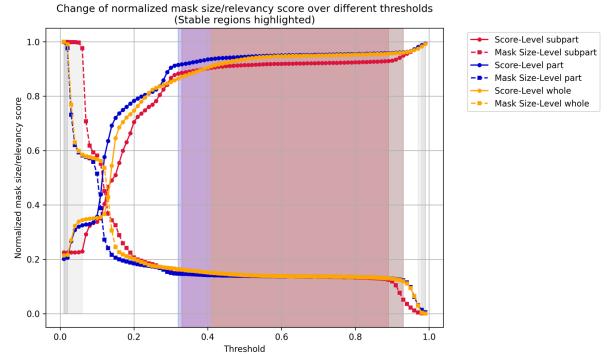


Figure 9. Relationship between score thresholds, normalized mask sizes, and relevancy scores for the "weaving basket" query in room scene. The plot shows how mask sizes decrease (downward trends with dotted lines) and average relevancy scores change (upward trends with smooth lines) as the threshold increases. Different hierarchical levels are represented by distinct colors, with the shaded region indicating the optimal stable region identified by our stability threshold.

7. More Visualizations

We show more examples on the 3D-OVS dataset in Figure 10, 11, and 12.

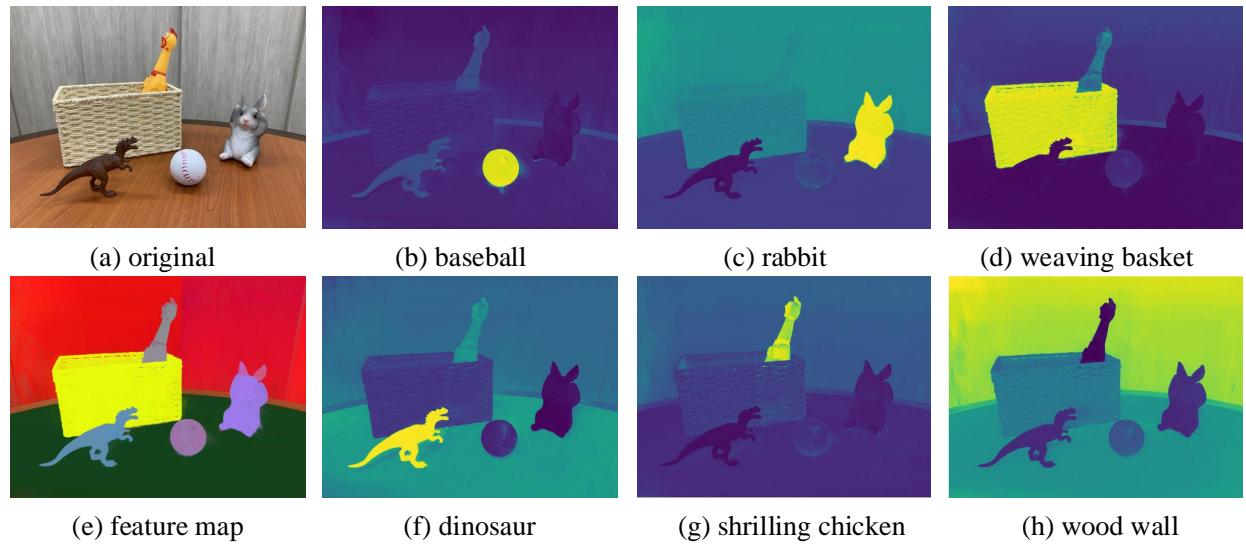


Figure 10. 3D-OVS: room

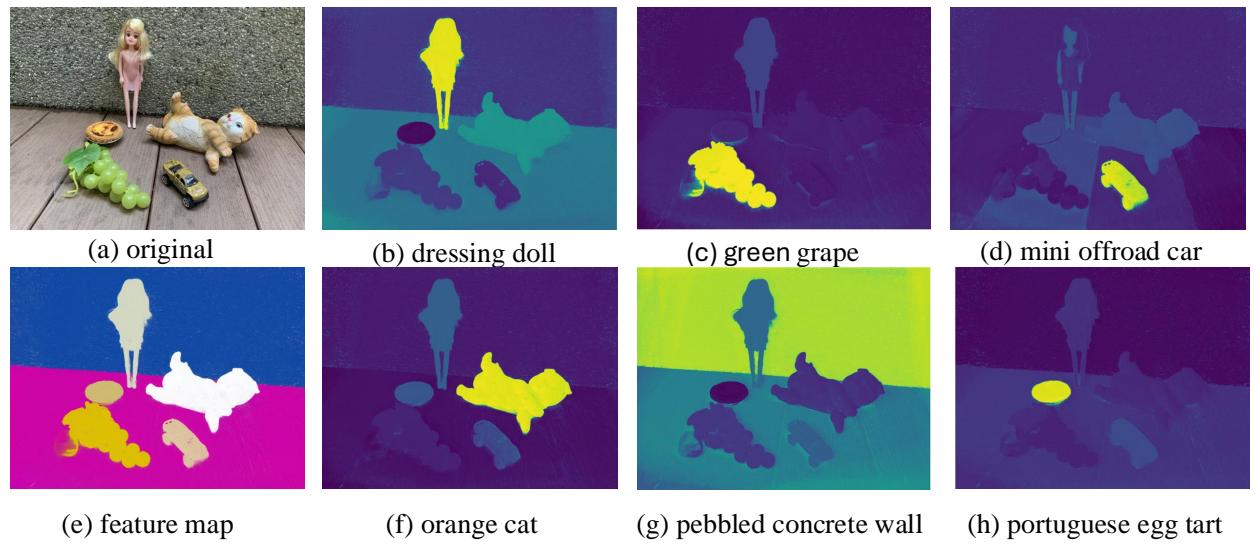


Figure 11. 3D-OVS: bench

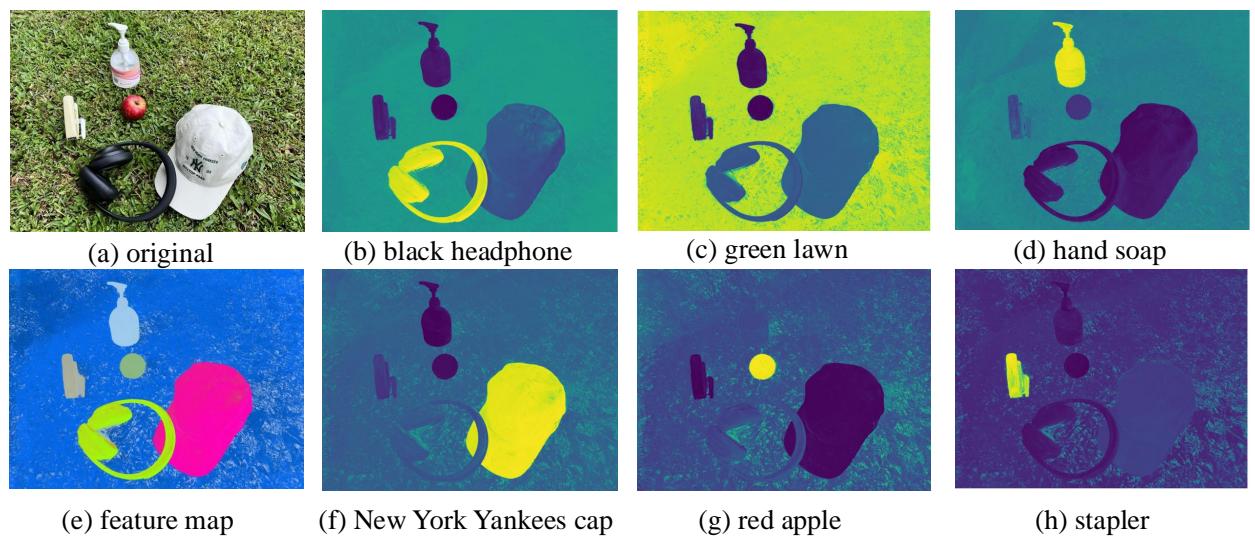


Figure 12. 3D-OVS: lawn