

SemanticSplat: Feed-Forward 3D Scene Understanding with Language-Aware Gaussian Fields

Qijing Li*

Jingxiang Sun*

Liang An

Zhaoqi Su

Hongwen Zhang

Yebin Liu[†]

Tsinghua University

Beijing Normal University

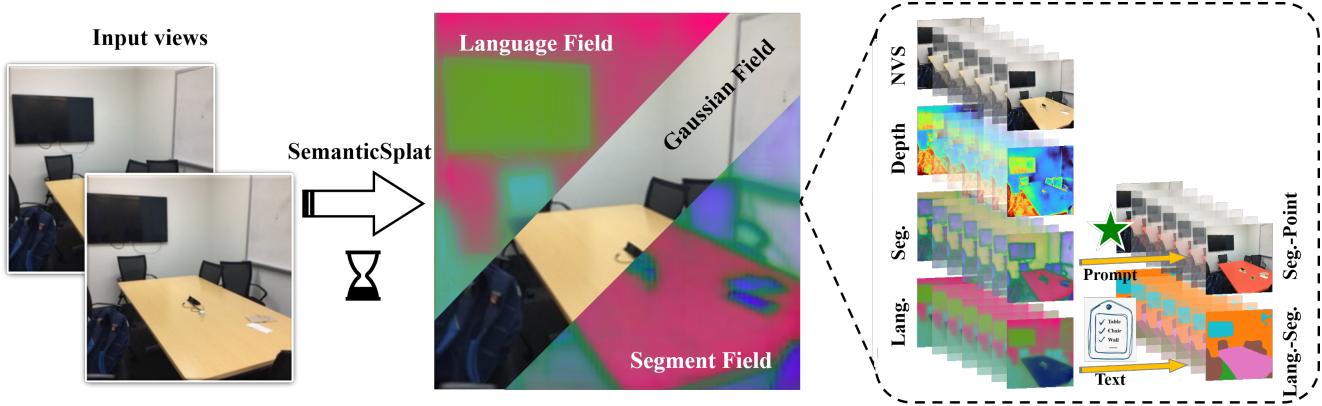


Figure 1. Our approach utilizes sparse view images as input to reconstruct a holistic semantic Gaussian field, which includes both the Gaussian field with language features and the segmentation features. This reconstruction captures geometry, appearance, and multi-modal semantics, enabling us to perform multiple tasks such as novel view synthesis, depth prediction, open-vocabulary segmentation, and promptable segmentation.

Abstract

Holistic 3D scene understanding, which jointly models geometry, appearance, and semantics, is crucial for applications like augmented reality and robotic interaction. Existing feed-forward 3D scene understanding methods (e.g., LSM) are limited to extracting language-based semantics from scenes, failing to achieve holistic scene comprehension. Additionally, they suffer from low-quality geometry reconstruction and noisy artifacts. In contrast, per-scene optimization methods rely on dense input views, which reduces practicality and increases complexity during deployment. In this paper, we propose SemanticSplat, a feed-forward semantic-aware 3D reconstruction method, which unifies 3D Gaussians with latent semantic attributes for joint geometry-appearance-semantics modeling. To predict the semantic anisotropic Gaussians, SemanticSplat fuses diverse feature fields (e.g., LSeg, SAM) with a cost volume representation that stores cross-view feature similarities, enhancing coherent and accurate scene comprehension. Leveraging a two-

stage distillation framework, SemanticSplat reconstructs a holistic multi-modal semantic feature field from sparse-view images. Experiments demonstrate the effectiveness of our method for 3D scene understanding tasks like promptable and open-vocabulary segmentation. Video results are available at <https://semanticsplat.github.io>.

1. Introduction

The ability to achieve holistic 3D understanding from 2D imagery is central to many applications in robotics, augmented reality (AR), and interactive 3D content creation. Such tasks demand representations that seamlessly combine precise geometry, realistic appearance, and flexible semantics. Traditional pipelines typically decompose this goal into multiple distinct stages: Structure-from-Motion (SfM) for sparse camera pose estimation, Multi-View Stereo (MVS) for dense geometry recovery, and specialized modules for semantic labeling. Although effective in structured scenarios, this staged approach is prone to error propagation—small inaccuracies

in early stages (e.g., pose estimation) often amplify through subsequent steps, resulting in degraded semantic and geometric reconstructions. Moreover, reliance on dense, accurately calibrated views severely restricts applicability in less controlled, real-world environments. Additionally, the lack of extensive labeled 3D datasets limits these methods’ ability to generalize beyond fixed semantic categories, hampering open-vocabulary scene understanding.

Recently, 3D scene understanding methods leveraging powerful pre-trained 2D foundational models—such as the Segment Anything Model (SAM) [20] and CLIP [33]—has emerged as a paradigm to enrich 3D representations with semantic knowledge distilled from readily available 2D data. However, directly transferring 2D semantic knowledge to 3D is non-trivial: 2D predictions often suffer from view-dependent inconsistencies, leading to noisy and unreliable semantic fields when aggregated across views. Besides, while NeRF [29] and explicit 3D Gaussian splatting [15] methods enhanced with 2D features have shown promise for open-vocabulary flexibility, existing approaches predominantly depend on per-scene optimization, making them impractical for dynamic or large-scale applications.

In this paper, we propose SemanticSplat, a feed-forward framework for joint 3D reconstruction and semantic field prediction from sparse input images. Our approach extends 3D Gaussian Splatting by augmenting each Gaussian with latent semantic attributes, enabling simultaneous rendering of RGB and semantic feature maps. This unified representation jointly encodes geometry and semantics within a single framework, where the learned semantic attributes maintain multi-view consistency while preserving the efficiency of 3D Gaussian representations. By distilling knowledge from pre-trained visual foundation models (VFM)s like SAM and CLIP-LSeg, we achieve robust and accurate promptable and open-vocabulary segmentation. Our key contributions include:

- 1. Feed-forward Holistic 3D Scene Understanding** – We propose a feed-forward semantic-aware method to predict semantic anisotropic Gaussians augmented with latent semantic features, enabling joint optimization of geometry, appearance, and multi-modal semantics. This facilitates a comprehensive understanding of 3D scenes.
- 2. Multi-Conditioned Feature Fusion** – We propose a novel pipeline that aggregates monocular semantic features (from SAM and CLIP-LSeg) with multi-view cost volumes, improving cross-view consistency and semantic awareness in complex scenarios.
- 3. Two-Stage Feature Distillation** – We separately lift SAM and CLIP-LSeg features into 3D through a two-stage process, reconstructing both segmentation and language feature fields. This supports multi-modal segmentation, including promptable and open-vocabulary segmentation.

2. Related Work

2.1. 3D Scene Understanding

Early language-aware scene representations embed CLIP features in NeRF volumes to support open-vocabulary queries, as demonstrated by LERF [16]. Subsequent efforts migrate to 3D Gaussian Splatting (3DGS) for real-time rendering: GARField [18] distills SAM masks into Gaussians, LangSplat [32] auto-encodes a scene-wise language field, and Gaussian Grouping [48] attaches identity codes for instance-level clustering. Recent extensions such as SAGA [2] and 4D LangSplat [24] provide promptable segmentation and temporally coherent language fields, respectively, yet their reliance on point-cloud surfaces limits mesh fidelity. Emerging approaches like OV-NeRF [25] introduce cross-view self-enhancement strategies to mitigate CLIP’s view inconsistency through semantic field distillation, while DiCo-NeRF [7] leverages CLIP similarity maps for dynamic object handling in driving scenes. Concurrent work MaskField [13] demonstrates how decomposing SAM mask features from CLIP semantics enables efficient 3D segmentation in Gaussian Splatting representation. Although fast in per-scene training, it still needs per-scene optimization.

2.2. Feed-Forward Gaussian Splatting

Feed-forward reconstructors amortize 3DGS inference. For example, PixelSplat [3] learns Gaussians from two views, while Splatter Image [41] accelerates single-view object recovery through per-pixel Gaussian prediction. The recent Hierarchical Splatter Image extension [36] introduces parent-child Gaussian structures to recover occluded geometry through view-conditioned MLPs. To leverage multi-view cues, MVSplat [6] builds cost volumes; we push this idea further by injecting monocular depth priors for texture-less scenes. Large-scale models trade hand-crafted geometry for data-driven priors: LGM [42], GRM [46], GS-LRM [49], and LaRA [4] reconstruct scenes in milliseconds but demand ≈ 60 GPU-days for pre-training. Gamba [37] achieves $1000\times$ speedup over optimization methods through Mamba-based sequential prediction of 3D Gaussians, though constrained to object-level reconstruction. Our approach reaches comparable quality in two GPU-days and, unlike LRMs, can be pre-trained with inexpensive posed images *without* depth supervision.

2.3. Lifting 2D Foundation Models to 3D

Neural fields can aggregate multi-view image features into a canonical 3D space. Semantic NeRF [51] and Panoptic Lifting [40] fuse segmentation logits, showing that consistent 3D fusion cleans noisy 2D labels. Beyond labels, Distilled Feature Fields [38], LERF [17], NeRF-SOS [11], and FeatureNeRF [47] render pixel-aligned DINO or CLIP em-

beddings for tasks such as key-point transfer. Recent 3DGS adaptations [12, 21, 23, 27, 32, 39, 53, 54, 56] adopt similar strategies to distill information from well-trained 2D models to 3D Gaussians. Feature 3DGS [53] generalizes distillation to explicit Gaussians; concurrent works like FMGS [56] and SPLAT-Raj [27] confirm that SAM or LSeg signals can be attached to Gaussians for open-vocabulary editing.

3. Method

Overview. The goal of our SemanticSplat is to holistically reconstruct the 3D scene with multi-modal semantics. As shown in Figure 2, given N sparse input images $\{I_i \in \mathbb{R}^{H \times W \times 3}\}_{i=1}^N$, with associated camera projection matrices $\{P_i = K_i [R_i | T_i]\}_{i=1}^N$ we propose to predict per-pixel semantic anisotropic Gaussians $\{(\mu_j, \alpha_j, \sum_j, c_j, f_j)\}_{j=1}^{H \times W \times N}$ for each image, representing the holistic semantic features of the scene, including segmentation features and language-aligned features. This enables feed-forward novel view synthesis and multi-modal segmentation of the scene, including promptable segmentation and open-vocabulary segmentation.

The per-pixel Gaussians are predicted through ViT-based feature matching using cost volumes, with per-view depth maps regressed by a 2D U-Net (Sec. 3.1). Inspired by Depth-Splat, we propose a new branch that conditions on multi-source monocular semantic features (Sec. 3.2) to enhance comprehension quality. In parallel with depth prediction, we introduce an auxiliary head to predict per-pixel semantic feature embeddings in 3D Gaussian space (Sec. 3.3). Leveraging a two-stage feature distillation process, we reconstruct a holistic semantic field lifted from 2D pretrained models (Sec. 3.4 and Sec. 3.5).

3.1. Efficient Depth Map Prediction

The first step of our SemanticSplat is to predict the depth map from the given inputs for further initiating the Gaussians. Typical image-to-image depth estimation pipelines like ViT-based encoder-decoder [10] often lead to noisy edge artifacts in rendering results, caused by directly regressing point maps from image pairs. Therefore, we instead estimate depth maps for both target and source RGB images through feature matching with cost volumes. This method aggregates feature similarities across views, thereby enhancing the model’s cross-view awareness. These depth maps are then converted to point clouds, upon which Gaussian parameters are regressed.

Multi-View Feature Extraction. We employ a CNN to extract down-sampled features from input views. These features are then processed by a Swin-Transformer [28, 44, 45], equipped with cross-attention layers to propagate informa-

tion across views, enhancing the model’s ability to capture inter-view relationships. The resulting multi-view-aware features are represented as $\left\{F_i \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C}\right\}_{i=1}^N$, where s is the down sampling factor and C is the feature dimension. Cross-attention is applied bidirectionally across all views, generalizing the framework to arbitrary numbers of input images.

Feature Matching and Depth Regression. Following MV-Splat [6], we adopt a plane-sweep stereo [8, 45] approach for cross-view feature matching. For each view i , we uniformly sample D depth candidates $\{d_m\}_{m=1}^D$ from the near-to-far depth range. Features from another view j are warped to view i at each depth candidate d_m , generating D warped features $\{F_{d_m}^{j \rightarrow i}\}_{m=1}^D$. The correlation between these warped features and view i ’s original features is computed to construct the cost volume $\{C_i \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times D}\}$. Finally, a 2D U-Net with a softmax layer predicts the per-view depth map by processing the concatenated Transformer features and cost volumes.

3.2. Multi-cond Semantic Features Aggregation

Recent Works [5] investigate the *geometry awareness* and *texture awareness* of visual foundation models (VFM) [1], which can enhance scene understanding. Leveraging the capabilities of VFM, we aggregate pre-trained monocular multi-task semantic features into the cost volume to address challenging scenarios.

Multi-conditioned Semantic Feature Fusion. We leverage the pre-trained segmentation backbone from the Segment Anything Model (SAM) [20] and CLIP-LSeg model [22] to get monocular features for each view. By interpolating to align with the cost volume resolution (Sec. 3.1), the processed segmentation-semantic features $\{F_i^{SAM} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C_{SAM}}\}_{i=1}^N$ from SAM and language-semantic features $\{F_i^{LSeg} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C_{LSeg}}\}_{i=1}^N$ from CLIP-LSeg are concatenated with cost volumes $\{C_i \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times D}\}_{i=1}^N$ and processed by a lightweight 2D U-Net [34, 35] to regress a unified latent feature map, integrating geometric and semantic cues.

3.3. Semantic Anisotropic Gaussians Prediction

Despite significant progress in scene understanding and language-guided reconstruction, existing methods [12, 23, 32, 51] often exhibit limited holistic scene comprehension and rely on single-modal segmentation. To address this, we propose holistic semantic field reconstruction via disentangled segmentation feature distillation and language feature

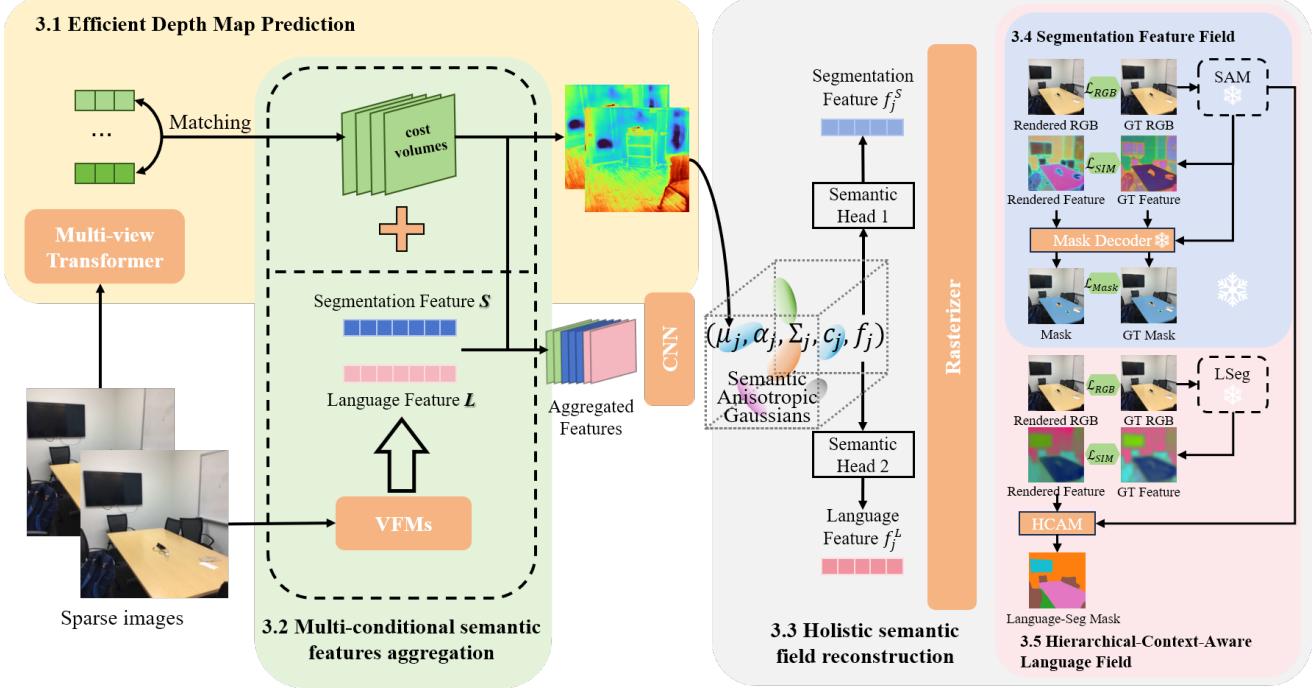


Figure 2. We employ multiview transformers with cross-attention to extract features from multi-view images and use cost volumes for feature matching (see Sec. 3.1). Utilizing the multi-conditioned semantic features from Visual Feature Modules (VFM’s) aggregated with the cost volumes (see Sec. 3.2), we predict Semantic Anisotropic Gaussians (see Sec. 3.3). Through a two-stage feature distillation process involving both segmentation (see Sec. 3.4) and language features (see Sec. 3.5), we reconstruct the holistic semantic feature field by jointly enforcing photometric fidelity and semantic consistency.

distillation, implemented through anisotropic semantic Gaussians.

Compared to conventional Gaussians $\{(\mu_j, \alpha_j, \sum_j, c_j)\}_{j=1}^n$ [15], semantic anisotropic Gaussians $\{(\mu_j, \alpha_j, \sum_j, c_j, f_j)\}_{j=1}^n$ incorporate latent space \mathbf{f} into Gaussian attributes to represent 3D semantic fields, enabling joint rendering of novel-view RGB map \mathbf{C} and semantic feature map \mathbf{F} .

$$\begin{cases} \mathbf{C} = \sum_{i=1}^n \mathbf{c}_i \alpha_i G_i(X) \prod_{j=1}^{i-1} (1 - \alpha_j G_j(X)), \\ \mathbf{F} = \sum_{i=1}^n \mathbf{f}_i \alpha_i G_i(X) \prod_{j=1}^{i-1} (1 - \alpha_j G_j(X)), \end{cases} \quad (1)$$

Here $G(X)$ stands for the projected 2D Gaussian kernel evaluated at pixel X .

Our pipeline proceeds as follows:

1. Initialization: Per-view depth maps from cost volumes are unprojected to 3D point clouds (as Gaussian centers μ_j) using camera parameters.
2. Attribute Prediction: Standard Gaussian parameters (opacity α_j , covariance Σ_j , color c_j) are predicted via two convolutional layers processing concatenated inputs (image features, cost volumes, and multi-view images). And the semantic latent attribute f_j is regressed from different

heads with the input as the latent feature map, depth, and view-aware features to match multi-modal segmentation.

Pre-trained visual foundation models (VFM’s) [1] often yield feature maps lacking view consistency and spatial awareness. To address this, we introduce a two-stage semantic feature distillation framework, including Segmentation Feature Field Distillation (Sec. 3.4) and Hierarchical-Context-Aware Language Field Distillation (Sec. 3.5) that lifts 2D features to 3D and refines latent feature maps, integrating diverse feature fields for holistic scene understanding.

3.4 Segmentation Feature Field

We leverage the Segment Anything Model (SAM) [20]—an advanced promptable segmentation model supporting inputs like points and bounding boxes—to distill segmentation-semantic embeddings into anisotropic Gaussians.

Feature Alignment. From the Gaussian semantic latent attribute f_j , an additional segmentation-semantic head is introduced to predict the segmentation-semantic f_j^S , as shown in Figure 2. After rasterization to 2D, we minimize the cosine similarity between the rasterized segmentation feature maps $S = \left\{ S_i \in \mathbb{R}^{h' \times w' \times d'} \right\}_{i=1}^N$, and the SAM encoder outputs

$\hat{S} = \left\{ \hat{S}_i \in \mathbb{R}^{H' \times W' \times C'} \right\}_{i=1}^N$. To improve efficiency and reduce memory consumption, we obtain compressed feature maps that are then upsampled to match the SAM features using CNN.

$$L_{dist}^{Seg} = 1 - \mathbf{sim}(f_{expand}(S), \hat{S}) = 1 - \frac{S \cdot \hat{S}}{\|S\| \cdot \|\hat{S}\|} \quad (2)$$

Prompt-Aware Mask Refinement. We integrate SAM’s pre-trained mask decoder into our pipeline to generate segmentation masks. A consistency loss enforces alignment between masks derived from our image embeddings $M = \left\{ M_i \in \mathbb{R}^{H \times W} \right\}_{i=1}^M$ and SAM embeddings $\hat{M} = \left\{ \hat{M}_i \in \mathbb{R}^{H \times W} \right\}_{i=1}^M$, ensuring promptable segmentation compatibility. We employ a linear combination of Focal Loss [26] and Dice Loss [30] in a 20:1 ratio for optimization.

Focal Loss:

$$FL(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (3)$$

Dice Loss:

$$d = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \quad (4)$$

$$L_{mask}^{Seg} = FL(p_t) + \frac{1}{20}d \quad (5)$$

3.5. Hierarchical-Context-Aware Language Field

We leverage CLIP-LSeg—a language-driven segmentation model that aligns textual descriptions with visual content via CLIP embeddings—to distill language-semantic embeddings into anisotropic Gaussians.

Feature Alignment. With the segmentation feature branch (Sec. 3.4) frozen for stability, we use a new language-semantic head to predict the language-semantic features f_j^L from the Gaussian semantic latent attribute f_j , as shown in Figure 2, and then rasterize it into 2D language feature maps $L = \left\{ L_i \in \mathbb{R}^{h'' \times w'' \times d''} \right\}_{i=1}^N$ and expand it to the dimension of CLIP-LSeg [22] feature maps $\hat{L} = \left\{ \hat{L}_i \in \mathbb{R}^{h'' \times w'' \times C''} \right\}_{i=1}^N$ as \bar{L} for loss computation.

$$L_{dist}^{Lang} = 1 - \mathbf{sim}(\bar{L}, \hat{L}) = 1 - \frac{\bar{L} \cdot \hat{L}}{\|\bar{L}\| \cdot \|\hat{L}\|} \quad (6)$$

Hierarchical-Context-Aware Pooling. We employ hierarchical-mask pooling on our expanded language-semantic features \bar{L} to enable fine-grained segmentation (e.g., object parts, materials): here we use SAM to extract

three-scale masks $\left\{ \mathbb{M}^h = \left\{ m_j^h \right\}_{j=1}^K \right\}_{h=s,m,l}$ (small/medium/large) to capture hierarchical object contexts. For each scale, L-Seg features within SAM-generated masks are aggregated via average pooling, enhancing intra-mask semantic consistency.

$$\bar{L}^h = \frac{\sum \bar{L} \cdot M^h}{\sum M^h} \quad (7)$$

3.6. Training Loss

During training, our model optimizes 3D anisotropic Gaussians $\{(\mu_j, \alpha_j, \Sigma_j, c_j, f_j)\}_{j=1}^{H \times W \times N}$ through a *two-stage loss formulation* with a combination of photometric loss and feature distillation loss that jointly enforces photometric fidelity and semantic consistency.

Photometric Loss.

$$\mathcal{L}_{rgb} = \sum_i \|\mathcal{R}_C(\mu, \alpha, \Sigma, c, f)_i - I_i^{\text{gt}}\|_1 + \lambda_1 \cdot \text{LPIPS}(\mathcal{R}_i, I_i^{\text{gt}}) \quad (8)$$

where \mathcal{R}_C is the differentiable renderer of image and I^{gt} the target image. And the loss weights of LPIPS [50] loss weight λ_1 is set to 0.05.

Semantic Distillation Loss (SAM Alignment):

$$\mathcal{L}_{sam} = L_{dist}^{Seg} + \lambda_{mask} L_{mask}^{Seg} \quad (9)$$

where λ_{mask} is set to 0.2.

Hierarchical-Context-Aware Distillation Loss (CLIP-LSeg Alignment, with segmentation feature branch frozen):

$$\mathcal{L}_{clip} = L_{dist}^{Lang} \quad (10)$$

4. Experiments

4.1. Settings

Datasets. We utilize the ScanNet [9] dataset for training, which provides high-fidelity 3D geometry and high-resolution RGB images, along with estimated camera intrinsic and extrinsic parameters for each frame. A total of 1,462 scenes are used for training, while 50 unseen scenes are reserved for validation. All frames are cropped and resized to a resolution of 256×256 .

Metrics. To evaluate photometric fidelity, we adopt standard image quality metrics: pixel-level PSNR, patch-level SSIM [43], and feature-level LPIPS [50]. For semantic segmentation, we measure performance using mean Intersection-over-Union (mIoU) and mean pixel accuracy (mAcc).



Figure 3. **Novel View Synthesis Comparisons.** Our method outperforms LSM and Feature-3DGS in challenging regions and is compatible with baseline MVSplat, which shows we reconstruct the appearance successfully

Table 1. **Comparison (Language-Seg).** Performance metrics for source and target view segmentation across different methods.

Method	Source View		Target View					
	mIoU↑	Acc.↑	mIoU↑	Acc.↑	PSNR↑	SSIM↑	LPIPS↓	
MVSplat	-	-	-	-	23.87	0.820	0.201	
LSeg	0.365	0.694	0.364	0.693	-	-	-	
LSM	0.347	0.679	0.347	0.679	17.84	0.630	0.372	
Feature-3DGS	0.510	0.804	0.235	0.585	13.00	0.407	0.600	
Ours	0.376	0.707	0.371	0.702	21.88	0.879	0.191	
Ours w/HCAM	0.376	0.707	0.386	0.710	21.88	0.879	0.191	

Implementation details Our model is trained using Adam [19] optimizer with an initial learning rate of $1e - 4$ and cosine decay following. Both semantic feature distillation stages are trained on 4 Nvidia A100 GPU for 5000 iterations.

4.2. Holistic Semantic Field Reconstruction

We compare our approach with two state-of-the-art methods: LSM [12] (a generalizable framework) and Feature-3DGS [52] (a per-scene optimization-based method). Both methods predict RGB values and leverage feature-based

Table 2. **Comparison (Promtable-Seg).** Performance metrics for source and target view segmentation across different methods.

Method	Source View		Target View					
	mIoU↑	Acc.↑	mIoU↑	Acc.↑	PSNR↑	SSIM↑	LPIPS↓	
SAM	0.684	0.427	0.426	0.684	-	-	-	
Feature-3DGS	0.678	0.409	0.395	0.681	13.14	0.405	0.601	
Ours	0.691	0.438	0.433	0.690	21.88	0.879	0.191	

3D Gaussian Splatting (3D-GS) [15]. Unlike our approach, Feature-3DGS supports promptable segmentation and open-vocabulary segmentation by separately optimizing SAM and LSeg features, while LSM is limited to open-vocabulary segmentation.

Evaluation of Novel View Synthesis We further compare our method with the baseline MVSplat [6], a feed-forward Gaussian reconstruction model trained on the RealEstate10K [55] dataset. As shown in Table 1 and Figure 3, Feature-3DGS [52] struggles to synthesize high-quality images from sparse input views, while LSM [12] introduces noisy artifacts, especially near object boundaries. Our

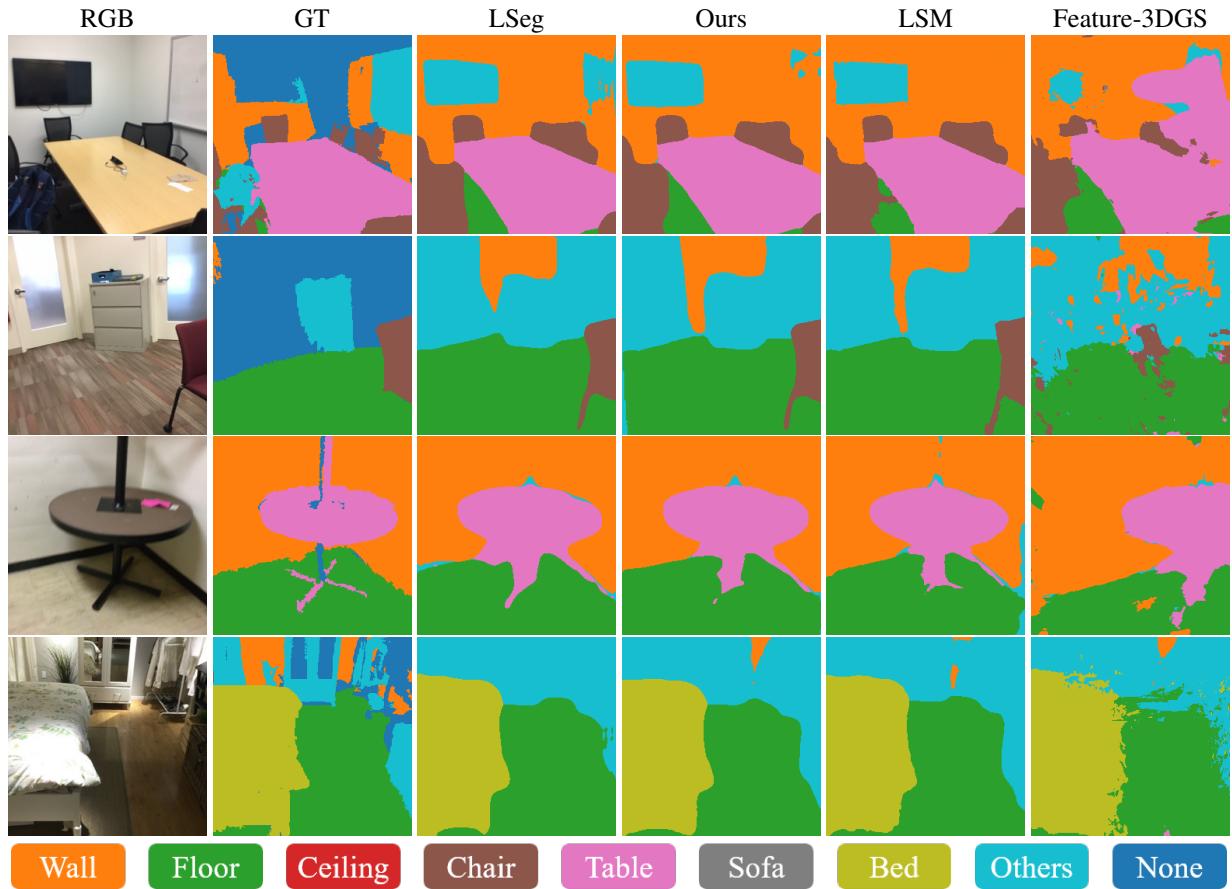


Figure 4. Language-based Segmentation Comparison. We visualize the segmentation from a set of categories for unseen view, our method outperforms with the other 3D method and comparably to the 2D VFM, which indicates we effectively lift 2d foundation language-image model to 3D.

method achieves comparable pixel-level quality to MVSplat despite training on lower-quality data and outperforms it at the patch and feature levels, owing to the semantic priors integrated into our pipeline.

Evaluation of Open-vocabulary Semantic 3D Segmentation

Following Feature-3DGS [52], we map thousands of category labels from diverse datasets into a unified set of common categories: {Wall, Floor, Ceiling, Chair, Table, Bed, Sofa, Others}. For comparison, we also include the 2D open-vocabulary segmentation method LSeg [22]. As shown in Table 1 and visualised in Figure 4, our method achieves competitive performance against baseline 3D methods and matches the accuracy of 2D methods when evaluated on the ScanNet dataset with transferred labels. Notably, while LSeg suffers from cross-view inconsistency, our approach maintains high consistency across views. To illustrate this, we visualize the language feature fields of both methods using PCA (projecting high-dimensional features into three channels) [14] in Figure 5.

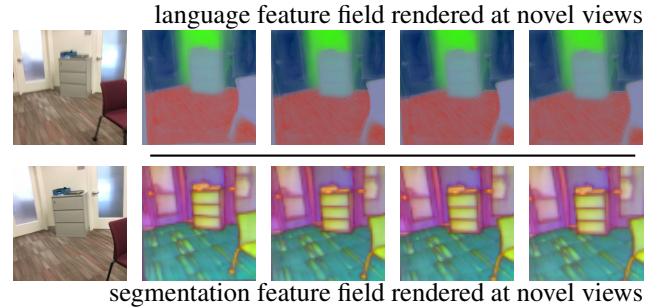


Figure 5. Visualization of the Semantic Feature Field. We visualize the language features and segmentation characteristics of the novel views, demonstrating how we elevate the 2D features into 3D while maintaining consistency across views. The visualizations are generated using PCA [31].

Evaluation of Promptable Semantic 3D Segmentation

Building on the SAM mask decoder, our method predicts three hierarchical masks from point queries and the predicted

Table 3. **Feature-Condition Ablation (Stages 1 & 2).** Stage 1: feature branch under LSeg vs. GT masks; Stage 2: feature branch under SAM vs. GT masks.

Condition	Compared with LSeg Masks		Compared with GT Masks		Compared with GT SAM		Compared with GT Masks	
	mIoU↑	Acc.↑	mIoU↑	Acc.↑	mIoU↑	Acc.↑	mIoU↑	Acc.↑
full	0.630	0.894	0.368	0.701	0.668	0.847	0.433	0.690
SAM	0.458	0.746	0.328	0.650	0.663	0.844	0.433	0.690
LSeg	0.628	0.893	0.369	0.699	0.591	0.790	0.424	0.681
w/o cond.	0.263	0.575	0.191	0.491	0.546	0.753	0.414	0.676

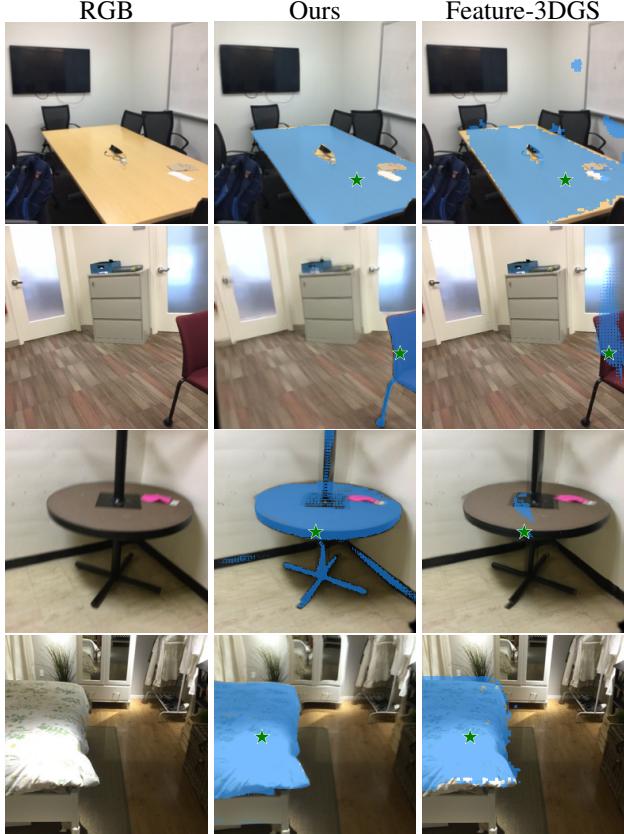


Figure 6. **Prompt-based Segmentation Comparison.** We visualize the segmentation generated from a point prompt for unseen views. Our method outperforms other 3D approaches, indicating that we effectively extend a 2D foundation segmentation model to 3D.

segmentation feature map. We uniformly sample a grid of points on the images (32 points along the width and 32 along the height, totaling 1,024 points). For each point, we generate hierarchical masks and evaluate their alignment with ground truth masks from ScanNet labels by reporting the highest Intersection-over-Union (IoU) and accuracy (Acc) scores in Table 2 and visualize the promptable segmentation in Figure 6. In addition to Feature-3DGS, we include the 2D promptable segmenter SAM in our comparisons. To visualize

the segmentation feature field and SAM features, we project them into three channels using PCA [14] in Figure 5.

4.3. Ablation Studies

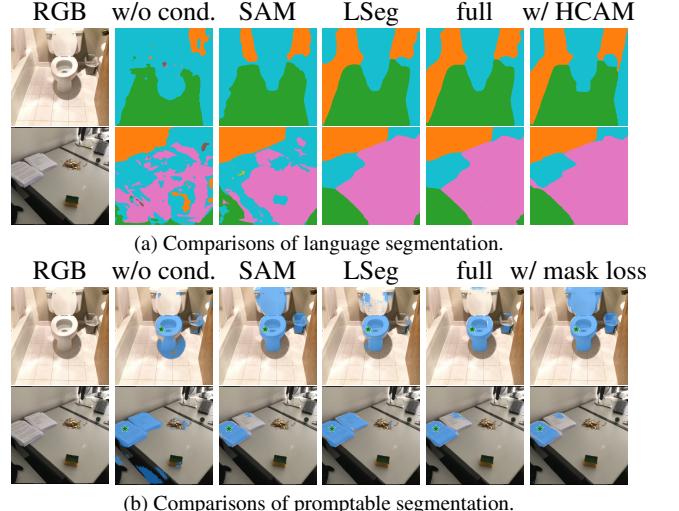


Figure 7. **Ablation study on different Conditions, HCAM and Mask loss.** We visualize the segmentation results under different conditions, illustrating that all these are complementary.

Multi-Conditioned Semantic Features In Table 3 and Figure 7, we compare our full model with a variant that excludes multi-semantic feature conditioning (Sec. 3.2), denoted as w/o condition, retaining only the multi-view branches. We further evaluate distinct monocular semantic features: the segmentation-semantic SAM feature and the language-semantic LSeg feature. To assess distillation and segmentation performance, we compare results separately using the 2D feature-derived masks and the ground truth (GT) masks. Our results demonstrate that feature concatenation achieves an optimal balance between promptable and open-vocabulary segmentation performance.

For the ablation study on Mask Loss in Semantic Distillation (Sec. 3.4), please refer to the supplementary material for more details.

5. Discussion

Limitation. While our method significantly reconstructs the holistic Gaussian feature field, it relies on a pre-trained model for feature lifting, which increases computational and GPU memory requirements. Additionally, our current model requires camera poses as input along with the multi-view images, which could limit scalability for various applications. Future work could explore pose-free models to move the requirement, further bridging the gap between modular design and real-world applicability.

Conclusion. In this paper, we propose a novel framework for feature distillation and multi-modal segmentation, leveraging multi-semantic conditioning with Segment Anything Model (SAM) and Language-Semantic (LSeg) features. Our experiments demonstrate that the complete model, incorporating both segmentation-semantic (SAM) and language-semantic (LSeg) features, achieves superior performance in balancing promptable and open-vocabulary segmentation tasks. In conclusion, this work advances the integration of vision-language models into 3D segmentation pipelines, offering a scalable solution for diverse semantic understanding tasks.

References

- [1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [2] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1971–1979, 2025.
- [3] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024.
- [4] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision*, pages 338–355. Springer, 2024.
- [5] Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. *arXiv preprint arXiv:2412.09606*, 2024.
- [6] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *ECCV*, 2024.
- [7] Jiho Choi, Gyutae Hwang, and Sang Jun Lee. Dico-nerf: Difference of cosine similarity for neural rendering of fisheye driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7858, 2024.
- [8] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition*, pages 358–363. Ieee, 1996.
- [9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [11] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776*, 2022.
- [12] Zhiwen Fan, Jian Zhang, Wenyang Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *NeurIPS*, 2024.
- [13] Zihan Gao, Lingling Li, Licheng Jiao, Fang Liu, Xu Liu, Wengping Ma, Yuwei Guo, and Shuyuan Yang. Fast and efficient: Mask neural fields for 3d scene segmentation. *arXiv preprint arXiv:2407.01220*, 2024.
- [14] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- [15] 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.
- [16] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023.
- [17] Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023.
- [18] Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21530–21539, 2024.
- [19] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [20] 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.

- [21] Hyunjee Lee, Youngsik Yun, Jeongmin Bae, Seoha Kim, and Youngjung Uh. Rethinking open-vocabulary segmentation of radiance fields in 3d space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4491–4498, 2025.
- [22] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.
- [23] Hao Li, Roy Qin, Zhengyu Zou, Diqi He, Bohan Li, Bingquan Dai, Dingewu Zhang, and Junwei Han. Langsurf: Language-embedded surface gaussians for 3d scene understanding. *arXiv preprint arXiv:2412.17635*, 2024.
- [24] Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, and Hanspeter Pfister. 4d langsplat: 4d language gaussian splatting via multimodal large language models. *arXiv preprint arXiv:2503.10437*, 2025.
- [25] Guibiao Liao, Kaichen Zhou, Zhenyu Bao, Kanglin Liu, and Qing Li. Ov-nerf: Open-vocabulary neural radiance fields with vision and language foundation models for 3d semantic understanding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [27] Xinyi Liu, Tianyi Zhang, Matthew Johnson-Roberson, and Weiming Zhi. Splatraj: Camera trajectory generation with semantic gaussian splatting. *arXiv preprint arXiv:2410.06014*, 2024.
- [28] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.
- [29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.
- [31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [32] 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.
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, pages 234–241. Springer, 2015.
- [36] Jianghao Shen, Nan Xue, and Tianfu Wu. A pixel is worth more than one 3d gaussians in single-view 3d reconstruction. *arXiv preprint arXiv:2405.20310*, 2024.
- [37] QiuHong Shen, Zike Wu, Xuanyu Yi, Pan Zhou, Hanwang Zhang, Shuicheng Yan, and Xinchao Wang. Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. *arXiv preprint arXiv:2403.18795*, 2024.
- [38] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *7th Annual Conference on Robot Learning*, 2023.
- [39] 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.
- [40] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023.
- [41] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10208–10217, 2024.
- [42] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.
- [43] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [44] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.
- [45] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13941–13958, 2023.
- [46] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d recon-

- struction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024.
- [47] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8962–8973, 2023.
- [48] 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.
- [49] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.
- [50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [51] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021.
- [52] 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.
- [53] 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.
- [54] Shijie Zhou, Hui Ren, Yijia Weng, Shuwang Zhang, Zhen Wang, Dejia Xu, Zhiwen Fan, Suya You, Zhangyang Wang, Leonidas Guibas, et al. Feature4x: Bridging any monocular video to 4d agentic ai with versatile gaussian feature fields. *arXiv preprint arXiv:2503.20776*, 2025.
- [55] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.
- [56] Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *arXiv preprint arXiv:2401.01970*, 2024.

Table 4. **Ablation Study.** Impact of mask loss on segmentation.

Variant	mIoU↑	Acc.↑
w/ mask loss	0.668	0.847
w/o mask loss	0.659	0.842

Table 5. **Inference Time per Module.**

Module	Time (s)
Depth map predict w/ condition (3.1+3.2)	0.092
Segmentation field distillation (stage 1)	0.032
Language field distillation (stage 2)	0.029
Total	0.153

A. Ablation on Mask Loss

We compare our first-stage feature distillation (full) with a variant that excludes the mask loss (w/o mask loss). We evaluate the segmentation metrics using SAM masks to assess the effectiveness of the distillation process in Table 4.

B. Module Timing

We evaluate the computational cost of each module by running inference on the ScanNet dataset and calculating the runtime for each component of our method, as detailed in Table 5.