Consistent Scene Understanding in 3D Gaussian Splatting via Multi-Cue Mask Refinement

Hyunjoon Park and Donghyeon Cho*

Department of Computer Science, Hanyang University, Seoul, South Korea
{junippini83, doncho}@hanyang.ac.kr

Abstract. Reliable instance-level scene understanding is a fundamental prerequisite for object-level interactions and high-fidelity 3D representations. While current methods often leverage 2D foundation segmentation models to obtain these priors, their 2D-centric design typically yields fragmented masks and inconsistent predictions across different views. To address these issues, we propose a novel framework that produces consistent 2D instance masks to guide the optimization of 3D Gaussian Splatting (3DGS) feature fields. Our framework consists of three main stages. (1) Multi-Cue Extraction that generates synergistic semantic, geometric, and structural priors from input images. (2) Multi-Cue-Guided Mask Merging process that consolidates fragmented masks using a composite merge score derived from semantic, depth, and edge cues. (3) Cross-View Mask Matching that establishes globally consistent identity assignments across all viewpoints. By transforming viewpoint-specific segments into coherent 3D primitives, our approach enables stable 3D instance segmentation and effective downstream editing tasks. Experiments demonstrate that our method significantly improves cross-view consistency and segmentation stability over existing baselines while maintaining high-fidelity photometric reconstruction.

Keywords: 3DGS · Over-segmentation · Multi-Cue Mask Refinement

1 Introduction

While Neural Radiance Fields (NeRF) [22] pioneer high-fidelity scene representation through continuous volumetric functions, 3D Gaussian Splatting (3DGS) [7] shifts the paradigm toward explicit primitive-based representations. Unlike implicit weight-based encoding of NeRF, 3DGS utilizes a discrete collection of Gaussians as a structured substrate, facilitating seamless integration with 2D foundation models like Segment Anything Model (SAM) [10], GroundingDINO [18] and Contrastive Language-Image Pretraining (CLIP) [26]. However, integrating these models into 3DGS remains challenging because 2D models lack 3D spatial awareness and geometric context and often produce inconsistent semantics

* Corresponding author.

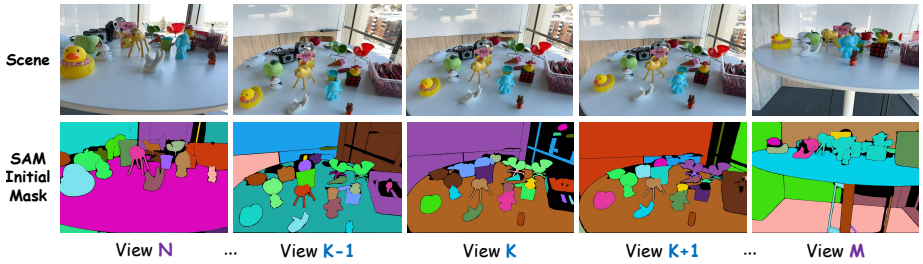

Fig. 1: **Cross-view inconsistency from SAM over-segmentation.** Automatic masks from SAM [10] suffer from inconsistent identities across distant $(N \dots M)$ and even contiguous frames $(K \pm 1)$. Its 2D-centric nature induces over-segmentation and erroneous merging of distinct objects, impeding coherent 3D scene understanding.

across views, making the transition from fragmented 2D segmentations to coherent 3D primitives unstable. Among these models, SAM is a representative 2D segmentation prior that is widely adopted for 3DGS-based scene editing. It offers dense, class-agnostic masks with precise boundaries for identifying editable object instances, which directly affects the 3D reasoning quality. Despite these advantages, the automatic mask generation of SAM often suffers from severe over-segmentation, partitioning a single object into numerous disconnected fragments (Fig. 1). This fragmentation stems from its 2D-centric design, which lacks depth awareness and cross-view semantic consistency. Consequently, visually similar but disjoint regions are erroneously merged, while minor appearance variations trigger unnecessary splits. Such inaccuracies hinder the establishment of stable object identities and lead to global ID inconsistencies. To address these challenges, we reinterpret the SAM output as an initially overcomplete decomposition that is subsequently refined using multiple cues. Specifically, we leverage three complementary cues: (1) DINOv2 [23] embeddings, which offer strong object-level semantics; (2) Monocular depth maps [32], which provides geometric context that is absent from purely appearance-based segmentation; and (3) LoG-filtered edge maps, which localize salient structural boundaries. Based on these cues, our model performs Multi-Cue-Guided Mask Merging (MCM) to mitigate the over-segmentation produced by SAM. In this process, depth acts as a hard constraint to prevent erroneous merges across disjoint boundaries, while semantic and structural affinities unify texture-induced splits. This resolves the "*white-on-white dilemma*", preserving distinct surfaces at varying depths while merging fragments. Subsequently, globally consistent identities established via cross-view mask matching are lifted into 3D using a robust multiview consensus procedure utilizing majority voting. Finally, joint optimization balances photometric fidelity with semantic regularization, yielding structured 3D masks as "*actionable primitives*" for high-fidelity scene understanding. Our contributions are as follows:

- Multi-Cue Mask Refinement framework that mitigates SAM-induced over-segmentation based on multiple cues.
- Cross-view mask matching that establishes globally consistent object IDs and suppresses view-dependent label inconsistencies across all viewpoints.
- Feature lifting that transfers 2D identities to 3D Gaussians while filtering unreliable assignments through majority voting and variance analysis.
- Extensive evaluation on LERF [8], Replica [28] and several real-world scenes, demonstrating substantial gains in correct object assignment and a significant reduction in over-segmented object counts.

2 Related Work

2.1 3D Feature Fields and Semantic Distillation

3D feature fields [11] have evolved from NeRF-based distillation [3,12,29] to real-time 3DGS frameworks [4,38] that embed 2D foundation priors (e.g., CLIP, DINOv2) into Gaussian primitives. Recent extensions incorporate hierarchical decomposition [1], part-level lifting [17,20], and 3D-aware prior fine-tuning [35]. While some approaches focus on multi-resolution alignment [8,25] or scalability through hash encodings [27,39], others improve multiview aggregation for static retrieval [14,30,31]. Unlike these methods, which treat feature fields as static tools, we actively reshape the segmentation by jointly reasoning over semantics, depth, and edges during optimization.

2.2 Object-Level Scene Understanding and Consistency

Bridging the gap between 2D segmentation and 3D object representation remains a fundamental challenge. Baselines such as GaussianGrouping [33], SAGA [2], and Splat-SAM [19] supervise identity encodings using SAM [10] masks. To improve consistency and boundary precision, subsequent works employ joint semantic optimization [16,34,37] or hierarchical clustering [9,13,6]. Despite these efforts, most methods inherit over-segmentation artifacts of SAM. In contrast, we treat SAM outputs as an initial over-complete decomposition, refining them through Multi-Cue-Guided Mask Merging (MCM) (Sec. 4.2) and cross-view mask matching (Sec. 4.3) to establish globally coherent 3D instances.

3 Preliminary

3.1 3D Gaussian Splatting (3DGS)

3D Gaussian Splatting (3DGS) [7] is an explicit scene representation that utilizes differentiable 3D Gaussian primitives. Each Gaussian is defined by its mean position $\boldsymbol{\mu}$, opacity α, and a 3D covariance matrix $\boldsymbol{\Sigma}$ derived from scaling $\mathbf{s}$ and rotation $\mathbf{q}$. The pixel color C is computed through tile-based rasterization and point-based α-blending: $C = \sum_{i \in \mathcal{N}} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j)$, where α_i represents the effective opacity at the pixel location. This framework can be extended to store high-dimensional feature vectors, forming a 3D feature field for semantic tasks.

3.2 Segment Anything Model (SAM)

Segment Anything Model (SAM) [10] is a foundation model for class-agnostic, promptable image segmentation. For autonomous scene decomposition, its Automatic Mask Generation (AMG) mode produces an initial set of binary masks $\mathcal{M} = \{m_i\}_{i=1}^{K}$. While SAM demonstrates exceptional zero-shot generalization, the resulting AMG masks are often 2D-centric and fragmented, lacking cross-view consistency in complex 3D environments.

4 Method

To address over-segmentation and view inconsistency in 2D-to-3D lifting, we propose a three-stage refinement pipeline (Fig 2). Rather than replacing SAM, we leverage its strength by treating its output as an initial over-complete decomposition that is then refined through semantic, geometric, and structural constraints (Sec. 4.1). First, we perform Multi-Cue-Guided Mask Merging (MCM) (Sec. 4.2) by integrating DINOv2 [23], monocular depth [32], and LoG edge [21] cues. This stage consolidates fragmented per-view masks into coherent regions using a scoring framework that balances semantic affinity with geometric boundary preservation. Second, we establish global object identities through cross-view mask matching and feature lifting (Sec. 4.3). By utilizing a multiview consensus voting mechanism, we effectively bypass the instability of 2D mask matching and ensure robust identity assignment for each Gaussian. Finally, a joint optimization (Sec. 4.4) enforces semantic consistency across the feature field while maintaining the high-fidelity photometric representation inherent to 3DGS. The comprehensive flow of our pipeline is summarized in Algorithm 1.

4.1 Multi-Cue Feature Extraction

Semantic Embeddings. We leverage DINOv2 [23] to provide global context that transcends local pixel appearances. For each candidate mask m_i generated by SAM, where i denotes the mask index, we extract dense feature maps and derive a representative descriptor f_i by averaging the feature vectors across the mask area. This descriptor serves as a semantic signature for calculating the cosine similarity between candidate masks in (1), enabling the framework to group fragments of the same entity regardless of minor text variations.
Monocular Depth Estimation. We leverage DepthAnythingV2 [32] to estimate per-pixel relative depth $D \in \mathbb{R}^{H \times W}$, providing geometric context absent from appearance-only segmentation [10]. For each mask m_i, we compute its mean depth $\bar{d}_i$ and depth gradient magnitude along its boundary ∂m_i. Despite scale ambiguity, this relative depth effectively disambiguates disjoint regions that share similar textures. These geometric priors enforce a hard depth constraint and govern the structural penalty terms in (1).
Boundary Localization. We employ Laplacian-of-Gaussian (LoG) [21] filtering to produce an edge-strength map ∇_{edge} for structural boundary localization.

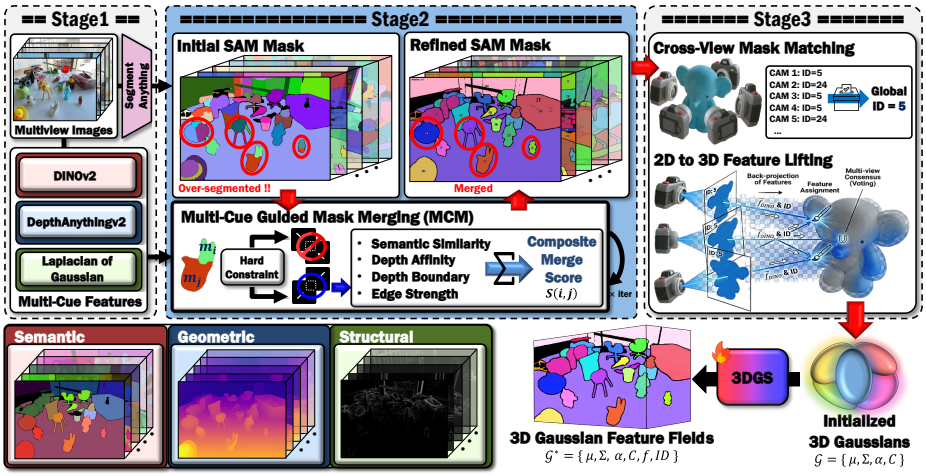

Fig. 2: **Pipeline**. Given multiview images, we extract initial SAM masks and multi-cue priors. These masks are refined via MCM using a composite merge score. Following cross-view mask matching, refined 2D features are back-projected and lifted into 3D Gaussians. Consequently, the model outputs a view-consistent object-level feature field for high-fidelity 3D scene understanding.

This rotation-invariant prior localizes structural discontinuities, ensuring precise alignment with object boundaries across diverse orientations. As a penalty term in (1), ∇_{edge} discourages the erroneous merging of distinct objects along their shared boundary ∂m_{ij}, even when they exhibit high semantic similarity.

4.2 Multi-Cue-Guided Mask Merging (MCM)

To resolve over-segmentation, we address the inherent limitations of semantic-only merging. These failures typically involve insufficient merging of noisy textures or erroneous grouping of visually similar objects located at distinct depths. To reduce these issues, we first utilize a hard depth constraint to exclude candidate pairs located at significantly different depths. Merge decisions for the remaining candidates are then decided by a composite merge score that synthesizes semantic and structural cues.

Adjacency Computation. To focus the merging process on plausible candidates, we define two masks m_i and m_j, where i and j denote the respective mask indices, as adjacent if they satisfy three geometric criteria: (1) Bounding box proximity within τ_{gap} pixels. (2) Intersection following morphological dilation such that $\text{dilate}(m_i) \cap \text{dilate}(m_j) \neq \emptyset$; and (3) Contact ratio exceeding τ_{contact}. This pruning strategy effectively eliminates over 80% of irrelevant candidate pairs, significantly reducing computational overhead while preserving all geometrically plausible merges.

Hard Depth Constraint. For these adjacent candidates, we first apply a hard geometric constraint to prevent erroneous merges across depth discontinuities by

enforcing $|\bar{d}_i - \bar{d}_j| < \tau_{\text{depth}}$ and depths are normalized to $[0, 1]$. This hard constraint ensures foreground-background separation even when semantic features align (e.g., identical textures at different depths), resolving the white-on-white dilemma where appearance-only methods fail.

Composite Merge Score. For adjacent pairs that satisfy the hard depth constraint, we compute a unified composite score to determine the final merge. This score synthesizes semantic similarity with structural and geometric penalties to ensure boundary-aware consolidation as follows:

$$\begin{aligned} S(i,j) = w_{\text{sem}} \cdot \cos(f_i, f_j) + w_{\text{depth}} \cdot \exp\left(-\frac{|\bar{d}_i - \bar{d}_j|^2}{2\sigma_d^2}\right) \\ - w_{\text{dbound}} \cdot \max(\nabla_D \cap \partial m_{ij}) - w_{\text{edge}} \cdot \max(\nabla_{\text{edge}} \cap \partial m_{ij}), \end{aligned} \tag{1}$$

where $S(i, j)$ evaluates the affinity between masks i and j. In this formulation, ∇_{edge} denotes the LoG edge magnitude, ∇_D represents the depth gradient, and $\partial m_{ij} = \partial m_i \cup \partial m_j$ represents the union of region boundaries. The composite merge score integrates four terms to balance semantic unification with structural preservation. First, the semantic term leverages high-dimensional DINOv2 embeddings to group fragments that belong to the same entity. Second, the depth affinity term promotes the unification of coplanar segments by evaluating their geometric continuity in 3D space. Third, the depth boundary term penalizes merges across sharp depth transitions to maintain separation between foreground and background objects. Finally, the edge magnitude term enforces structural integrity by preventing the consolidation of distinct objects separated by strong visual boundaries. Through this synthesis, $S(i, j)$ optimizes the trade-off between recall and precision during the process. The framework executes a merge only when a mask pair satisfies the hard depth constraint and exceeds the threshold τ_{merge}. Under this protocol, pairs failing the initial constraint are immediately classified as distinct entities and excluded from further evaluation. This ensures that only spatially plausible candidates proceed to the final assessment, where their structural and semantic coherence is verified for unification.

Iterative Refinement. We apply greedy merging iteratively (max 1,000 iterations), recomputing adjacency and scores after each merge. The process terminates when no pair satisfies all constraints.

4.3 Cross-View Mask Matching & Feature Lifting to 3D Gaussians

Cross-View Mask Matching. The preceding stage produces geometrically coherent but viewpoint-specific mask regions. To establish global consistency, we perform Global Mask Association by constructing an affinity matrix from the representative descriptors $\bar{f}_i$, representing the mean features of consolidated masks. Two masks m_i and m_j from different viewpoints are assigned a shared global identity if their cosine similarity exceeds the matching threshold τ_{match}:

$$\cos(\bar{f}_i, \bar{f}_j) > \tau_{\text{match}}. \tag{2}$$

Algorithm 1: Overview of Proposed Refinement Pipeline

Require: Images $\{I_n\}_{n=1}^{N}$, Initial Masks $\{M_n\}_{n=1}^{N}$, 3D Gaussians $\mathcal{G}$
Ensure: Feature-enriched 3D Gaussians $\mathcal{G}_{feat}$

Stage I: Multi-Cue Feature Construction // Sec. 4.1
for each view $n \in \{1, \ldots, N\}$ **do**
  Extract DINOv2 (f), LoG Edges (∇_{edge}), and Depth (D)
end for

Stage II: Multi-Cue-Guided Mask Merging // Sec. 4.2
for each view n **do**
  repeat
    Identify adjacent masks m_i, m_j
    if $|\bar{d}_i - \bar{d}_j| < \tau_{depth}$ **then**
      Compute composite merge score $S(i, j)$ via (1)
      if $S(i, j) > \tau_{merge}$ **then**
        Update $m_{new} \leftarrow \text{Merge}(m_i, m_j)$
      end if
    end if
  until convergence
end for

Stage III: Cross-view Mask Matching and 3D Lifting // Sec. 4.3
$\mathcal{C}_{global} \leftarrow \text{GlobalClustering}(\{m_n\})$ via DINO similarity
for each Gaussian $g \in \mathcal{G}$ **do**
  Project g onto visible camera views $n \in \mathcal{V}_g$
  $ID_g \leftarrow \text{MajorityVote}(\text{Global IDs from projected views})$
  $\sigma^2_{feat} \leftarrow \text{ComputeVariance}(g, \mathcal{V}_g)$
  if $\sigma^2_{feat} > \tau_{var}$ **then**
    Filter g (Reliability check)
  end if
end for

return Final optimized feature field $\mathcal{G}_{feat}$

This deterministic association enables the GlobalClustering of masks across the entire scene, ensuring that unique object instances maintain stable identities throughout the reconstructed scene.

Feature Lifting to 3D Gaussians. Once global identities are established across all viewpoints, we lift these 2D semantic assignments into a 3D Gaussian representation to form a consistent feature field. This process bridges the gap between viewpoint-specific 2D masks and the underlying 3D structure by populating each Gaussian g with a representative identity feature, f_g. Distinct from the 2D semantic descriptor f_i, the identity feature f_g serves as a high-dimensional 3D signature that anchors global object consistency across the Gaussian field.

Multi-View Semantic Fusion. For each 3D Gaussian g with center position μ_g, we aggregate information from all N cameras through a weighted fusion of the corresponding 2D features:

$$f_g = \frac{\sum_{v \in \mathcal{V}g} w_v \cdot f_{m_v(g)}}{\sum_{v \in \mathcal{V}g} w_v} \tag{3}$$

where $\mathcal{V}g$ is the set of viewpoints where Gaussian g is visible, aggregation weight w_v, and $f_{m_v(g)}$ is the refined identity feature of the mask $m_v(g)$. This weighted scheme prioritizes viewpoints with clear, orthogonal visibility, thereby minimizing the impact of perspective distortion and partial occlusions. The resulting feature f_g serves as a robust 3D descriptor. enabling the subsequent assignment of stable object identities through the multiview consensus mechanism.

3D Mask ID Assignment. To elucidate object representation, each Gaussian is assigned a definitive label using a multiview majority voting mechanism. The assigned identity ID_g is determined by identifying the most frequent global ID across all the visible viewpoints:

$$\mathrm{ID}_g = \arg\max_k \sum_{v \in \mathcal{V}_g} \mathbb{1}[\mathrm{ID}_{m_v(g)} = k], \tag{4}$$

where $\mathrm{ID}_{m_v(g)}$ is the global identity of the mask containing the projection of Gaussian g in view v, and $\mathbb{1}[\cdot]$ denotes the indicator function. This voting strategy effectively filters out transient single-view noise and misalignments, ensuring that each 3D primitive is assigned to the most statistically reliable region. By establishing a deterministic mapping between global IDs and their respective color encodings, this mechanism ensures that refined masks exhibit identical identity values and consistent color signatures across all viewpoints.

Variance-based Reliability Filtering. To ensure that only consistent multiview evidence is used, we perform a statistical validation of the lifted features. For each 3D Gaussian g, we compute the feature variance $\sigma^2_{\text{feat}}(f_g)$ across all views where the primitive is visible:

$$\sigma^2_{feat}(f_g) = \frac{1}{|\mathcal{V}_g|} \sum_{v \in \mathcal{V}_g} \|f_{m_v(g)} - f_g\|^2. \tag{5}$$

where $\mathcal{V}_g$ is the set of viewpoints in which gaussian g satisfies the visibility criteria, $|\mathcal{V}_g|$ is the total number of visible views, and $f_{m_v(g)}$ represents the 2D identity feature projected from view v. Gaussians with $\sigma^2_{feat}(f_g) > \tau_{\text{var}}$ are discarded as unreliable. Such high-variance assignments typically arise from occlusion boundaries or inconsistent 2D segmentation across different viewpoints. This filtering ensures that the noise inherent in the single-view model outputs is reduced.

4.4 Joint Optimization

The 3D representation is optimized end-to-end using a composite loss function:

$$\mathcal{L} = \mathcal{L}_{\text{render}} + \lambda_s \mathcal{L}_{\text{semantic}} + \lambda_{\text{reg}} \mathcal{L}_{\text{3D-reg}}, \tag{6}$$

where $\mathcal{L}_{\text{render}}$ represents the photometric rendering loss following standard 3DGS [7], $\mathcal{L}_{\text{semantic}}$ denotes the identity consistency loss, and $\mathcal{L}_{\text{3D-reg}}$ signifies the structural 3D regularization term, with λ_s and λ_{reg} serving as balancing weights.

Semantic Cohesion Loss. To ensure that the 3D representation inherits the refined 2D semantics, we attach learnable semantic features to each Gaussian and supervise them using the lifted features.

$$\mathcal{L}_{\text{semantic}} = \frac{1}{|\mathcal{G}_{\text{valid}}|} \sum_{g \in \mathcal{G}_{\text{valid}}} \|f_g - \tilde{f}_g\|^2, \tag{7}$$

where $\tilde{f}_g$ is the target lifted feature and $\mathcal{G}_{\text{valid}}$ denotes the subset of Gaussians with a reliable multiview consensus. Although semantic features are optimized jointly with geometry and appearance, this loss is applied only after the completion of the densification process to prevent semantic gradients from interfering with the initial geometric convergence.

3D Geometric Regularization. We encourage spatial smoothness via k-NN consistency [24] as follows:

$$\mathcal{L}_{\text{3D-reg}} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \sum_{g' \in \mathcal{N}_5(g)} \|f_g - f_{g'}\|^2, \tag{8}$$

where $\mathcal{N}_5(g)$ denotes 5 nearest spatial neighbors. This regularization produces coherent semantic fields that are suitable for object-level manipulation.

Two-Stage Optimization. We adopt two-stage training schedule: (1) Stage 1 (iter 0–15k): Standard 3DGS optimization focusing on geometry and appearance, with semantic features frozen. (2) Stage 2 (iter 15k–100k): Joint optimization of geometry, appearance, and semantic features with $\mathcal{L}_{\text{semantic}}$ and $\mathcal{L}_{\text{3D-reg}}$ activated. This schedule stabilizes early reconstruction while ensuring convergence to semantically coherent representations

5 Experiments

5.1 Experimental Settings

We implement our proposed framework based on the official codebase of GaussianGrouping [33]. All 3DGS [7] parameters and feature fields are optimized for 30,000 iterations on a single RTX 3090 GPU. For evaluation, we utilize the LERF [8], Replica [28], and in-the-wild datasets from [15] to validate robustness across diverse real-world environments. The rendering quality is assessed via PSNR, SSIM [5], and LPIPS [36] to ensure that our joint optimization preserves a high-fidelity photometric representation while regularizing the feature field. Segmentation performance is measured through mIoU and Mask mIoU (mMIoU), while Boundary mIoU (mBIoU) specifically evaluates mask alignment with physical boundaries for actionable primitives. Finally, we report the average mask count (#Mask) per scene to quantify the capacity of our framework to reduce over-segmentation in the raw foundation model outputs.

Hyperparameters. For mask merging in (1), we empirically set weights to $w_{sem} = 0.5$, $w_{edge} = 0.2$, $w_{depth} = 0.2$, and $w_{dbound} = 0.1$ with $\sigma_d^2 = 0.05$. To satisfy the dual condition for merging, we define two key thresholds: the merging

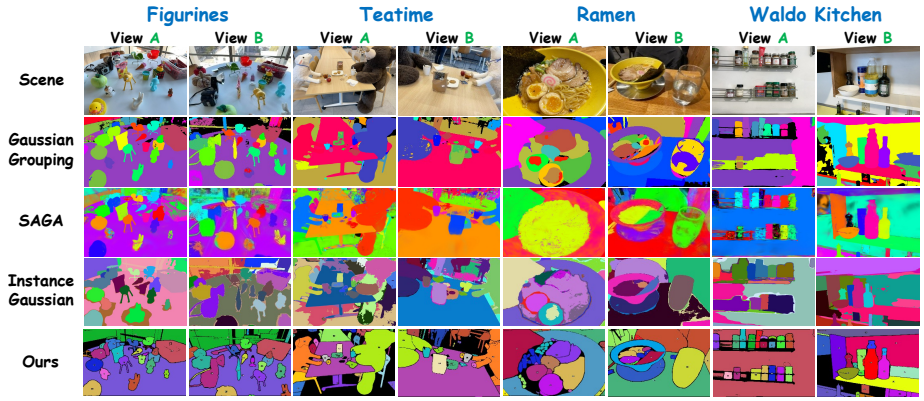

Fig. 3: **Qualitative Result.** This figure demonstrates how faithfully our results represent the scene and maintain view consistency across different viewpoints.

threshold $\tau_{merge} = 0.85$, and the hard constraint $\tau_{depth} = 0.15$. While τ_{depth} acts as a hard constraint to filter out disjoint regions, τ_{merge} determines the final merge based on the composite merge score. For cross-view matching, we set the similarity threshold $\tau_{match} = 0.7$, and a variance threshold $\tau_{var} = 0.05$ in (5). These values remain fixed across all datasets to demonstrate the robustness of our framework.

5.2 Qualitative Results

We qualitatively compare our method against state-of-the-art (SOTA) 3D segmentation frameworks [33,2,16]. As illustrated in Fig. 3, while baseline methods achieve reasonable object localization, they consistently struggle with identity fragmentation and boundary leakage. Specifically, GaussianGrouping tends to produce over-segmented results. This fragmentation occurs because these methods directly inherit the raw output of SAM without accounting for the underlying 3D geometric continuity. Furthermore, InstanceGaussian occasionally suffers from "*label bleeding*" where object boundaries become blurred or misaligned in complex scenes with overlapping depth layers. In contrast, our approach yields significantly cleaner and more stable object masks by employing MCM (Sec. 4.2). By evaluating fragmented regions through scoring mechanism, we successfully unify texture-induced fragments.

5.3 Quantitative Results

Table 1 summarizes the quantitative performance of our method compared to vanilla 3DGS [7], SOTA 3D segmentation [2,16,33] and feature field based approaches [9,14,38] on the LERF and Replica datasets.

Photometric Rendering Quality. As shown in Table 1, our method maintains high-fidelity rendering performance comparable to the vanilla 3DGS. While

Table 1: **Quantitative Comparison on LERF and Replica Dataset.** The **bold** score is the best, and the underline score is the second.

Category	Model	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	mBIoU ↑	#Mask ↓
Standard	3DGS [7]	28.6	0.88	**0.11**	-	-	-
3D Segmentation	Gau-Group [33]	28.6	0.88	0.12	0.685	0.642	9,627
	SAGA [2]	28.2	0.87	0.15	0.712	0.658	178
	InstanceGaussian [16]	**28.8**	0.87	0.15	0.702	0.662	101
Feature Field	Feature3DGS [38]	27.5	0.84	0.18	0.538	0.485	-
	GARField [9]	28.1	0.87	0.16	0.710	0.615	-
	CF3 [14]	28.3	0.87	0.14	0.524	0.582	-
Proposed	**Ours**	28.6	**0.89**	0.13	**0.728**	**0.677**	**67**

InstanceGaussian achieves the highest PSNR of 28.8, our framework matches the baseline at 28.6 and achieves a competitive SSIM of 0.89. This indicates that our joint optimization and mask merging process does not compromise the underlying photometric reconstruction quality of the 3D Gaussian representations.

Segmentation Accuracy. In terms of 3D segmentation, our method achieves SOTA performance across all primary metrics. We report a Mean IoU (mIoU) of 0.728, outperforming the second-best method, SAGA [2] (0.712), and significantly surpassing feature field-based approaches like Feature3DGS [38] (0.538). Furthermore, our method excels in boundary precision, reaching a Mean Boundary IoU (mBIoU) of 0.677. This improvement highlights the effectiveness of our multi-cue integration, particularly the use of LoG-based structural priors and depth-aware constraints, in producing sharp and physically accurate object boundaries compared with existing distilled feature field methods.

Resolution of Over-segmentation. The most significant contribution of our framework is the drastic reduction in the over-segmentation inherent in foundation model priors. While GaussianGrouping generates an average of 9,627 masks per scene due to the fragmented nature of SAM outputs, our strategy successfully merges them into 67 masks. Even compared to other models like SAGA and InstanceGaussian, our method produces the most compact and semantically consistent object representations. This reduction confirms that our approach effectively extracts "*actionable*" object primitives suitable for downstream tasks.

5.4 Analysis

View Consistency. We qualitatively evaluate cross-view consistency using the "Ramen" scene. While the baseline [33] inherits over-segmentation of SAM and produces fluctuating identities across different views, our framework merges these fragments through Sec. 4.2. As illustrated in Fig. 4 (View A–E), our model maintains stable object identities and sharp boundaries despite significant viewpoint changes. These results demonstrate that our refined masks serve as robust

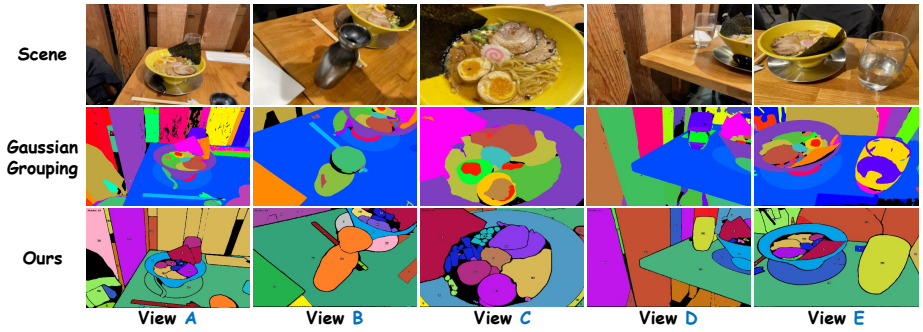

Fig. 4: **Qualitative Comparison of View Consistency.** Compared to the baseline [33], which exhibits over-segmentation and inconsistent object identities across different views, our approach successfully produces coherent, view-consistent object masks across changing viewpoints.

Table 2: **Ablation Studies on LERF Figurine Dataset.** This table quantifies the incremental performance gains achieved by sequentially integrating semantic [23], geometric [32], and structural [21] priors into our framework.

Configuration	Signals			Metrics				
	DINOv2	LoG	DAV2	mIoU ↑	# Masks ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline [33]	–	–	–	0.685	9,627	**28.6**	0.88	0.13
Ours (1)	✓	–	–	0.721	245	28.4	0.89	0.14
Ours (2)	✓	✓	–	0.729	112	28.5	**0.91**	**0.12**
Ours (Full)	✓	✓	✓	**0.741**	**67**	**28.6**	**0.91**	**0.12**

* DAV2 denotes DepthAnythingV2.

actionable primitives, successfully bridging the gap between fragmented 2D priors and coherent 3D object representations.

Impact of Multi-Cue Priors. Table 2 quantifies the incremental gains from integrating semantic, structural, and geometric priors. The baseline [33] exhibits extreme fragmentation with 9,627 masks, illustrating the pitfalls of lifting raw SAM outputs directly into 3D spaces. Adding DINOv2 semantic priors (Ours (1)) reduces the mask count to 245, proving that high-level context is essential for initial object consolidation. Incorporating structural priors via LoG (Ours (2)) further refines these groupings with sharp edge guidance and improves rendering metrics through precise boundary delineation. Our full model, which includes DepthAnythingV2 geometric priors, achieves state-of-the-art performance with 0.741 mIoU and a consolidated set of 67 actionable primitives.

Impact of Cross-View Mask Matching. To validate cross-view mask matching (Sec. 4.3), we compare our framework against a baseline that refines masks only within individual views. As shown in Fig. 5, per-view refinement resolves local over-segmentation but fails to maintain identity consistency across different

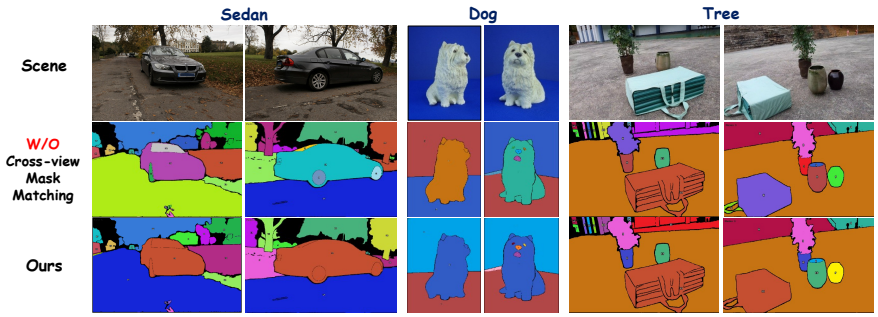

Fig. 5: **Qualitative Ablation of Cross-View Mask Matching.** Without cross-view matching (middle row), masks are restricted to per-view assignments, leading to inconsistent identities. Our full pipeline (bottom row) ensures global identity coherence, which is essential for stable 3D representation.

viewpoints. This lack of coherence results in identity flickering, where a single object is assigned different IDs across viewpoints. Such instability prevents our mechanism (Sec. 4.3) from effectively aggregating features into coherent 3D primitives. Our strategy resolves this issue by establishing stable correspondences through global semantic affinities. This approach ensures consistent object assignment and transforms the segments into coherent actionable primitives.

6 Conclusion

We present a unified framework for 3D scene decomposition that transforms fragmented 2D priors into structured and view-consistent object primitives. By treating the foundation model outputs as an initial overcomplete state, our approach effectively bridges the gap between per-view 2D perception and coherent 3D representation. The proposed Multi-Cue-Guided Mask Merging process addresses over-segmentation of SAM through multi-cue integration, while cross-view mask matching and feature lifting ensure global identity coherence. These components enable the extraction of high-fidelity primitives that maintain geometric and semantic consistency across viewpoints, providing a robust foundation for structured 3D scene understanding and downstream applications.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25422680, Metacognitive AGI Framework and its Applications, 33 %), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University), 33 %) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00521432, 34 %).

References

1. Bhalgat, Y., Laina, I., Henriques, J.F., Zisserman, A., Vedaldi, A.: N2f2: Hierarchical scene understanding with nested neural feature fields. In: ECCV. pp. 197–214. Springer (2024) 3
2. Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Segment any 3d gaussians. In: AAAI. vol. 39, pp. 1971–1979 (2025) 3, 10, 11
3. Chou, Z.T., Huang, S.Y., Liu, I., Wang, Y.C.F., et al.: Gsnerf: Generalizable semantic neural radiance fields with enhanced 3d scene understanding. In: CVPR. pp. 20806–20815 (2024) 3
4. Guo, J., Ma, X., Fan, Y., Liu, H., Li, Q.: Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. arXiv (2024) 3
5. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: ICPR. pp. 2366–2369. IEEE (2010) 9
6. Hu, X., Wang, Y., Fan, L., Luo, C., Fan, J., Lei, Z., Li, Q., Peng, J., Zhang, Z.: Sagd: Boundary-enhanced segment anything in 3d gaussian via gaussian decomposition (2025), https://arxiv.org/abs/2401.17857 3
7. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023) 1, 3, 8, 9, 10, 11
8. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: ICCV. pp. 19729–19739 (2023) 3, 9
9. Kim, C.M., Wu, M., Kerr, J., Goldberg, K., Tancik, M., Kanazawa, A.: Garfield: Group anything with radiance fields. In: CVPR. pp. 21530–21539 (2024) 3, 10, 11
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023) 1, 2, 3, 4
11. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. NeurIPS **35**, 23311–23330 (2022) 3
12. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. NeurIPS **35**, 23311–23330 (2022) 3
13. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: CVPR. pp. 4558–4567 (2018) 3
14. Lee, H., Min, J., Park, J.: Cf3: Compact and fast 3d feature fields. ICCV (2025) 3, 10, 11
15. Lee, Y., Ryu, J., Yoon, D., Cho, D.: Consistent object removal from masked neural radiance fields by estimating never-seen regions in all-views. In: ICPR. pp. 416–431. Springer (2024) 9
16. Li, H., Wu, Y., Meng, J., Gao, Q., Zhang, Z., Wang, R., Zhang, J.: Instance-gaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In: CVPR. pp. 14078–14088 (2025) 3, 10, 11
17. Liu, M., Uy, M.A., Xiang, D., Su, H., Fidler, S., Sharp, N., Gao, J.: Partfield: Learning 3d feature fields for part segmentation and beyond. In: ICCV. pp. 9704–9715 (2025) 3
18. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55. Springer (2024) 1
19. Lu, Y., Zhou, Y., Qiao, Y., Song, C., Liang, T., Ma, J., Wang, H., Yin, Y.: Segment then splat: Unified 3d open-vocabulary segmentation via gaussian splatting. In: NeurIPS 3

20. Ma, Z., Yue, Y., Gkioxari, G.: Find any part in 3d. In: ICCV. pp. 7818–7827 (2025) 3
21. Marr, D., Hildreth, E.: Theory of edge detection. Proceedings of the Royal Society of London. Series B. Biological Sciences **207**(1167), 187–217 (1980) 4, 12
22. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 1
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv (2023) 2, 4, 12
24. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009) 9
25. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: CVPR. pp. 20051–20060 (2024) 3
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021) 1
27. Shi, J.C., Wang, M., Duan, H.B., Guan, S.H.: Language embedded 3d gaussians for open-vocabulary scene understanding. In: CVPR. pp. 5333–5343 (2024) 3
28. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv (2019) 3, 9
29. Tschernezki, V., Laina, I., Larlus, D., Vedaldi, A.: Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In: 3DV. pp. 443–453. IEEE (2022) 3
30. Wang, N., Yan, X., Song, X., Wang, Z.: Semantic-guided gaussian splatting with deferred rendering. In: ICASSP. pp. 1–5. IEEE (2025) 3
31. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., et al.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. NeurIPS **37**, 19114–19138 (2024) 3
32. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. NeurIPS **37**, 21875–21911 (2024) 2, 4, 12
33. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: ECCV. pp. 162–179. Springer (2024) 3, 9, 10, 11, 12
34. Yin, Y., Liu, Y., Xiao, Y., Cohen-Or, D., Huang, J., Chen, B.: Sai3d: Segment any instance in 3d scenes. In: CVPR. pp. 3292–3302 (2024) 3
35. Yue, Y., Das, A., Engelmann, F., Tang, S., Lenssen, J.E.: Improving 2d feature representations by 3d-aware fine-tuning. In: ECCV. pp. 57–74. Springer (2024) 3
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018) 9
37. Zhang, W., Zhang, L., Hu, P., Ma, L., Zhuge, Y., Lu, H.: Bootstraping clustering of gaussians for view-consistent 3d scene understanding. In: AAAI. vol. 39, pp. 10166–10175 (2025) 3
38. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: CVPR. pp. 21676–21685 (2024) 3, 10, 11
39. Zuo, X., Samangouei, P., Zhou, Y., Di, Y., Li, M.: Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. IJCV **133**(2), 611–627 (2025) 3