Supplementary Material of "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

1 Summary of Supplementary Material

In this supplementary document, we first describe the implementation details (see Sec. 2). We then provide detailed explanations of our Model Selection, that is, why we choose DINOv2 [8], DepthAnythingv2 [15], and LoG [7] among several models (see Sec. 3). We then present additional qualitative results of mask refinement (see Sec. 4), and ablation studies of the hyperparameters utilized in our main algorithm (see Sec. 5) and metrics (see Sec. 6). Finally, we discuss the current limitations and potential directions for future work.

2 Implementation Details

2.1 Network Architecture

DINO Feature Extraction. We employ DINOv2 [8] (ViT-B/14) as the pre-trained backbone for semantic feature extraction, leveraging its discriminative representation for robust object-level understanding. Input images are resized to 518×518 pixels, ensuring precise alignment with the 14×14 patch size. The model yields 768-dimensional per-patch features from the final layer, capturing dense semantic signatures even in cluttered environments. To facilitate stable cosine similarity calculations, all extracted descriptors are L_2-normalized before the Multi-Cue-Guided Mask Merging (MCM) process.

Object Feature Field. Each 3D Gaussian is augmented with a learnable 16-dimensional object feature vector $\mathbf{o} \in \mathbb{R}^{16}$, which is initialized from a zero-centered Gaussian distribution $\mathcal{N}(0, 0.01)$. These features are jointly optimized with the Gaussian geometry and appearance parameters to enable consistent object-level grouping. During rendering, the object features are splatted via the standard 3DGS rasterization pipeline, producing 16-channel identity feature maps for 3D-to-2D supervision.

* Corresponding author.

Table 1: **Quantitative Analysis of Semantic Backbones.** We assess the discriminative power of various priors using Intra-object and Inter-object cosine similarities. The Separation Ratio (↑) represents the relative gap between internal consistency and external distinctness.

Feature	Intra ↑	Inter ↓	Ratio ↑	Dim
CLIP-ViT-B/16	0.71	0.43	1.65	512
SAM mask feat.	0.68	0.51	1.33	256
DINOv2-Small [8]	0.84	0.29	2.90	384
DINOv2-Base [8]	**0.89**	**0.21**	**4.24**	768

2.2 Optimization Details

Learning Rates. We employ the Adam optimizer for all parameters. The learning rates for each parameter type are as follows:

- **Position (xyz)**: Exponential decay from 1.6×10^{-4} to 1.6×10^{-6} over 50k iterations.
- **Object features**: 2.5×10^{-3} (fixed during Stage 1).
- **Opacity**: 5.0×10^{-2}.
- **Scaling**: 5.0×10^{-3}.
- **Rotation**: 1.0×10^{-3}.

Training Schedule. The total optimization process takes 100k iterations on a single NVIDIA RTX 3090 GPU. Densification and pruning are performed every 100 iterations between 500 and 15k iterations. To ensure that the semantic features do not destabilize the geometric convergence, we freeze the 16-dimensional object features during the first 15k iterations. After the densification phase, we activate the joint optimization with $\lambda_s = 0.1$ and $\lambda_{reg} = 0.01$.

3 Model Selection

Our framework integrates three distinct foundation models to achieve robust multi-cue feature extraction. Here, we provide an extended justification for these architectural decisions, supported by theoretical analysis and empirical observations.

3.1 Monocular Depth Estimation: DepthAnythingv2

We employ DepthAnythingv2 [15] as our primary geometric prior. Unlike earlier monocular depth estimators such as MiDaS [12] or DPT [11], which often produce over-smoothed depth maps, Depth Anything V2 exhibits an exceptional boundary sharpness. This is critical for our framework because our hard depth constraint relies on detecting sharp discontinuities to prevent "bleeding" between foreground objects and the background. Furthermore, training the model on a massive 62M

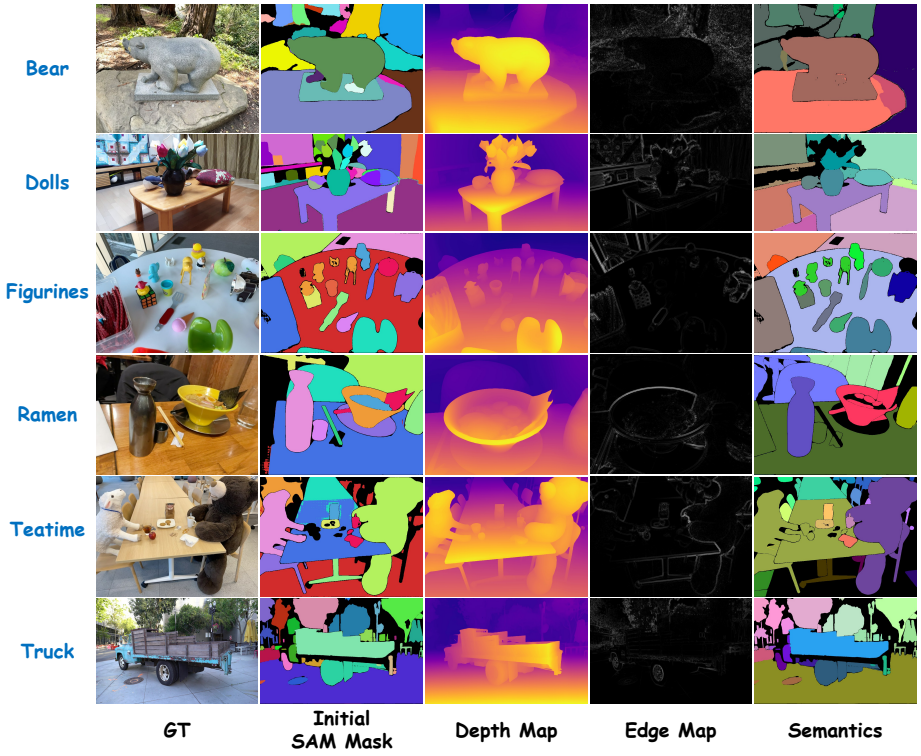

Fig. 1: **Visualization of multi-cue features.** We extract complementary priors to address the limitations of the 2D-centric proposals of SAM. While the Initial SAM Mask exhibits severe oversegmentation, our framework leverages Depth Maps (geometric context), Edge Maps (structural discontinuities), and Semantics (high-level object coherence) to consolidate fragmented regions. This multi-cue integration ensures that the resulting segments align precisely with physical object boundaries across diverse scenes

image dataset provides a level of zero-shot robustness that is essential for diverse 3DGS captures (e.g., Mip-NeRF 360 [1], LERF [3]). In our experiments, we observed that while appearance-only features (DINOv2) often suffer from visual aliasing, where two white walls at different depths produce identical semantic embeddings, Depth Anything V2 successfully resolves this ambiguity. By providing the missing geometric context, it allows the system to treat these walls as physically disjoint entities, effectively solving the "white-on-white dilemma."

3.2 Semantic Feature Extraction: DINOv2

The choice of DINOv2 [8] over language-aligned models such as CLIP [10] or LSeg [6] is driven by the requirement for spatially dense representations. CLIP

is optimized for global image-text alignment, resulting in coarse feature maps that lack the local geometric precision required for lifting features to individual 3D Gaussians. In contrast, DINOv2 provides patch-level features (vitb14, 768D) that preserve the local structure of scenes. While SAM features are excellent for boundary localization, they often lack the semantic depth required to associate fragments across multiple views. DINOv2's self-supervised pre-training enables it to capture fine-grained visual similarities (e.g., different parts of a complex chair) that are often absent in text-supervised models, providing a more stable foundation for our feature lifting and consensus voting stages.

3.3 Edge Detection: Laplacian-of-Gaussian (LoG)

For structural boundary detection, we utilize the Laplacian of Gaussian (LoG) [7] instead of modern learning-based detectors or multi-stage algorithms such as Canny [2]. A key requirement for our composite score $S(i, j)$ is a continuous edge response. While Canny employs non-maximum suppression and hysteresis thresholding to produce binary edges, LoG provides a continuous strength map. This allows our scoring mechanism to weigh edge proximity probabilistically rather than being forced into a binary decision that may introduce artifacts. Theoretically, the second-derivative formulation of the LoG is rotation invariant, making it superior to directional gradient operators (such as Sobel [13]) when processing complex 3D scenes with arbitrary edge orientations. We utilize a Gaussian kernel with $\sigma = 2.0$ to perform multi-scale filtering, which suppresses high-frequency texture noise (e.g., a patterned rug) while preserving salient structural boundaries of objects. This parameter-free approach ensures that our edge detection remains robust across different datasets without requiring the retraining or fine-tuning that learning-based detectors would necessitate.

3.4 Synthesis of Multi-Cues

The integration of these three models addresses the fundamental limitations of single-modality reasoning. While depth resolves appearance aliasing, semantics capture high-level coherence, and edge detection identifies boundaries that may be obscured by semantic similarities. Specifically, the complementary nature of these cues facilitates robust decision-making in complex environments. For instance, in texture-less regions where semantic embeddings (f_i) may appear near-identical, the monocular depth prior ($\bar{d}_i$) provides critical evidence for foreground-background separation via a hard depth constraint. Simultaneously, the LoG-based edge maps (∇_{edge}) act as structural stabilizers, preventing the overconsolidation of adjacent objects that share similar depth values but belong to distinct semantic categories. This tripartite synergy ensures that our "actionable primitives" are both geometrically accurate and semantically consistent across all views. By synthesizing these signals, the framework effectively resolves the "white-on-white dilemma" and preserves the fine-grained structural integrity that single-view 2D priors typically overlook.

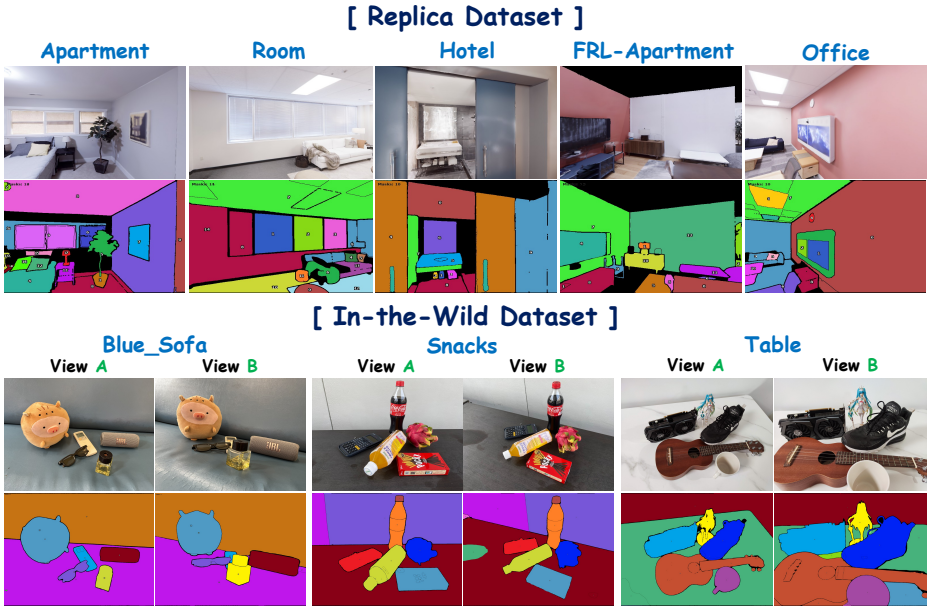

Fig. 2: **Additional Qualitative Result.** Our method successfully generates globally coherent 3D instance masks across various indoor and real-world scenes.

4 Additional Qualitative Results

We provide further qualitative evidence to demonstrate the robustness and generalization of our proposed framework across a variety of 3D environments, including synthetic indoor scenes and real-world captures.

Generalization on the Replica Dataset. The upper row of our qualitative gallery showcases the performance of our method on the Replica dataset [14], which comprises diverse indoor environments such as Apartment, Room, Hotel, FRL-Apartment, and Office. As illustrated in Fig. 2, our framework consistently produces semantically meaningful 3D instances across these varied layouts. Specifically, the model effectively isolates individual furniture items and architectural elements with high precision, even in cluttered office or residential settings. By leveraging the synergy between DINOv2-based semantic features and geometric priors, our approach successfully resolves the oversegmentation typical of 2D foundation models, ensuring that large surfaces, such as walls and floors, are consolidated into singular, coherent masks.

Consistency in In-the-Wild Scenarios. To verify the practical applicability of our method, we evaluate it on "In-the-Wild" datasets featuring real-world objects in uncontrolled lighting and complex backgrounds, as shown in the bottom panel. In the Blue_Sofa, Snacks, and Table sequences, we observe that our method maintains remarkable identity stability across significant viewpoint transitions (View A and View B). Unlike per-view segmentation baselines that suffer from

Table 2: **Sensitivity Analysis of Hyperparameters.** We evaluate the impact of merging thresholds (τ) and feature weights (w) on the Replica dataset. The **Ours** configuration achieves the optimal balance between object coherence and reconstruction fidelity.

Config.	τ_{merge}	τ_{depth}	w_{sem}	Rendering Metrics			Segmentation	
				PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	# Mask ↓
Aggressive Merging	0.70	0.30	0.5	27.8	0.88	0.15	0.624	**32**
Conservative Merging	0.95	0.05	0.5	**28.6**	0.90	0.13	0.691	412
Structural-biased	0.85	0.15	0.2	27.9	0.89	0.14	0.652	89
Semantic-biased	0.85	0.15	0.8	28.3	0.90	0.13	0.685	124
Ours	**0.85**	**0.15**	**0.5**	**28.6**	**0.91**	**0.12**	**0.728**	67

flickering labels, our framework ensures that each object—ranging from intricate snack packaging to complex arrangements of items on a table—retains a consistent identity and sharp boundary across different camera angles. These results confirm that our joint reasoning over semantic, depth, and edge cues provides a robust foundation for stable object-level scene understanding in diverse real-world configurations.

5 Ablation Studies

Table 2 presents a comprehensive sensitivity analysis of our framework's key hyperparameters, including merging thresholds ($\tau_{merge}, \tau_{depth}$) and semantic feature weights (w_{sem}). We evaluate these configurations to identify the optimal balance between photometric fidelity and object coherence.

Impact of Merging Thresholds. The choice of thresholds significantly dictates the level of mask consolidation. Aggressive Merging ($\tau_{merge} = 0.70$) achieves the most compact representation with only 32 masks; however, it suffers from a substantial drop in segmentation accuracy (0.624 mIoU) and rendering quality (27.8 PSNR). This performance degradation stems from the overconsolidation of physically distinct objects that share similar semantic or geometric signatures. Conversely, Conservative Merging ($\tau_{merge} = 0.95$) maintains high photometric fidelity but fails to resolve the fundamental oversegmentation of SAM, resulting in an excessive mask count of 412.

Impact of Feature Weights. The balancing weight w_{sem} modulates the influence of the DINOv2 semantic embeddings relative to the geometric and structural cues. A structure-biased configuration ($w_{sem} = 0.2$) leads to fragmented results because it overlooks object-level semantic continuity. Conversely, a semantic-biased setup ($w_{sem} = 0.8$) tends to ignore geometric boundaries, causing "label bleeding" across depth discontinuities.

Optimal Configuration. Our default configuration (Ours) employs $\tau_{merge} = 0.85, \tau_{depth} = 0.15$, and $w_{sem} = 0.5$. This setup achieves a state-of-the-art balance, yielding the highest mIoU (0.728) and SSIM (0.91) while effectively consolidating

the scene into 67 actionable primitives. This confirms that our multi-cue scoring framework is the most robust when the semantic and geometric signals are weighted equally.

6 Evaluation Metrics

We employ multiple metrics to evaluate segmentation quality, cross-view consistency, and editing applicability.

Mask Intersection-over-Union (mIoU). The standard metric for segmentation quality is as follows: Given the predicted mask M_p and ground truth mask M_g:

$$\text{IoU}(M_p, M_g) = \frac{|M_p \cap M_g|}{|M_p \cup M_g|}. \tag{1}$$

For multi-object scenes, we compute mean IoU (mIoU) by matching predicted and ground truth masks using the Hungarian algorithm [5]:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}(M_p^i, M_g^{\pi(i)}), \tag{2}$$

where π is the optimal assignment, and N is the number of ground-truth objects.

Boundary mIoU To evaluate boundary quality, we compute IoU only within a narrow band around object boundaries. Let $\mathcal{B}_\epsilon(M)$ denote the ϵ-dilated boundary of mask M:

$$\mathcal{B}_\epsilon(M) = (M \oplus \epsilon) \setminus (M \ominus \epsilon), \tag{3}$$

where $\oplus$ and $\ominus$ denote dilation and erosion operations, respectively, with a disk kernel of radius $\epsilon = 5$ pixels. This metric is crucial for 3DGS-based editing, where sharp boundaries are required to prevent texture bleeding during the manipulation of objects. Boundary mIoU is:

$$\text{Boundary-mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|(M_p^i \cap M_g^{\pi(i)}) \cap \mathcal{B}_\epsilon(M_g^{\pi(i)})|}{|(M_p^i \cup M_g^{\pi(i)}) \cap \mathcal{B}_\epsilon(M_g^{\pi(i)})|}. \tag{4}$$

Mask Count (#Mask). To quantify the degree of over-segmentation, which is a primary limitation of raw SAM proposals, we recorded the total number of unique masks $|\mathcal{M}|$ generated for each scene. A lower #Mask value, when paired with a high mIoU, indicates that the framework successfully consolidates fragmented regions into coherent, semantically meaningful object instances.

7 Limitations and Future Works

7.1 Limitations.

Despite its improved coherence, our framework has two primary limitations. First, it assumes that SAM [4] provides an oversegmented initial state. Thus, severe undersegmentation or missed boundaries are irrecoverable through merging alone. Second, because per-view merging operates independently before global clustering, extreme occlusions or appearance changes may yield localized inconsistencies.

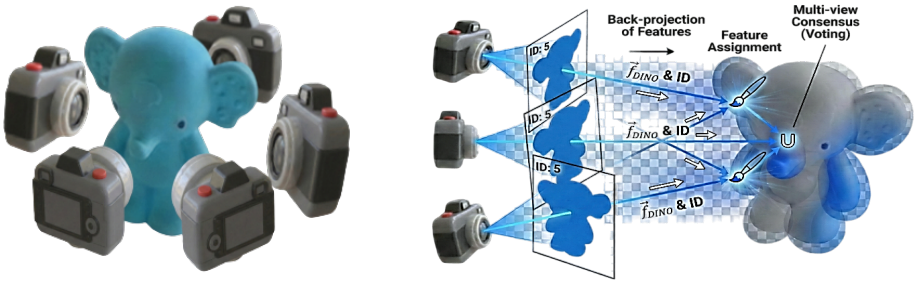

Fig. 3: **AI generated Images.** We utilized these images in our main pipeline to illustrate effectively.

7.2 Future Works.

Our pipeline significantly enhances index-based editing in frameworks such as GaussianGrouping [16] by providing unified "*actionable primitives*" and eliminating manual fragment selection. In the future, we aim to utilize these consistent 3D masks as robust conditions for generative tasks, such as high-fidelity object replacement, which currently suffer from boundary artifacts. Furthermore, extending our multi-cue reasoning to dynamic 3DGS could enable temporally stable tracking and semantic manipulation in complex spatiotemporal scenes. Finally, integrating language-guided models, such as CLIP [10] or LangSplat [9] with our refined feature fields will allow for intuitive text-driven object disambiguation, further bridging the gap between raw scene representation and user-centric 3D content creation.

8 Generative AI Disclosure.

The authors utilized the Gemini Pro model to assist in generating the illustrative components of the pipeline diagram, specifically for visualizing the camera perspectives that surround the target object (elephant) and the back-projection of features into 3D Gaussians ($G = \{\mu, \Sigma, \alpha, C\}$), as shown in Fig. 3. To ensure transparency, we provide the specific prompts used for the generation process:

- **Prompt 1 (Object Isolation):**
  "An illustration showing a target object (an elephant) isolated from a complex scene using a segmentation mask, with multiple camera frustums from different angles pointing at the object to represent a multiview capture setup."
- **Prompt 2 (Feature Lifting):**
  "A conceptual diagram showing the back-projection of 2D object features into a 3D space. The 2D masks with global IDs are lifted and integrated into 3D Gaussian primitives, illustrating the formation of a view-consistent 3D feature field."

The AI-generated components were used solely for the visual communication of technical concepts and did not contribute to the data analysis or the scientific results of this study.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV. pp. 5855–5864 (2021)
2. Canny, J.: A computational approach to edge detection. TPAMI (6), 679–698 (1986)
3. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: ICCV. pp. 19729–19739 (2023)
4. 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)
5. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
6. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation (2022), `https://arxiv.org/abs/2201.03546`
7. Marr, D., Hildreth, E.: Theory of edge detection. Proceedings of the Royal Society of London. Series B. Biological Sciences **207**(1167), 187–217 (1980)
8. 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 preprint arXiv:2304.07193 (2023)
9. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: CVPR. pp. 20051–20060 (2024)
10. 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)
11. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
12. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. TPAMI **44**(3), 1623–1637 (2020)
13. Sobel, I., Feldman, G.: An isotropic 3x3 gradient operator for image processing. Presentation at Stanford Artificial Intelligence Project pp. 271–272 (1968)
14. 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 preprint arXiv:1906.05797 (2019)
15. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. NIPS **37**, 21875–21911 (2024)
16. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: ECCV. pp. 162–179. Springer (2024)