

SLARM: Streaming and Language-Aligned Reconstruction Model for Dynamic Scenes

Supplementary Material

1. More Implementation Details

Model architecture. As the standard configuration, our model employs a 12-layer Alternating-Attention Transformer [8], which interleaves frame-wise and global self-attention mechanisms. Each attention layer operates with a feature dimensionality of 768. The input image is processed using a Vision Transformer (ViT) with a patch size of 8×8 , yielding a sequence of image tokens that serve as input to a Gaussian Decoder. This decoder comprises three lightweight MLP-based task-specific heads: a *Gaussian head*, a *motion head*, and a *semantic head*.

The Gaussian head regresses the geometric and appearance parameters of 3D Gaussians, specifically the pixel-aligned depth $d \in \mathbb{R}$, rotation represented by a unit quaternion $\mathbf{q} \in \mathbb{R}^4$, scale $\mathbf{s} \in \mathbb{R}^3$, opacity $\alpha \in [0, 1]$, and color $\mathbf{c} \in \mathbb{R}^3$, collectively forming a 12-dimensional parameter vector per Gaussian primitive.

Each Gaussian primitive is parameterized by position, scale, rotation (quaternion), opacity, RGB color, and an auxiliary depth value. We use the following activation functions to map raw network outputs to valid physical ranges:

- **Scale:** $\text{scale} = \min(\exp(x + \text{scale_offset}), 0.5)$, where $\text{scale_offset} = -0.693$ (i.e., $\log(0.5)$). This initialization biases the model to start from relatively large Gaussians and shrink during training, which we empirically find beneficial for stable self-supervised learning of motion.
- **Opacity:** $\sigma = \text{sigmoid}(x - 2.0)$, following GSLRM [9], which encourages sparse initialization (low opacity) and reduces floaters.
- **RGB:** $\mathbf{c} = \text{sigmoid}(x)$, clamping colors to $[0, 1]$.
- **Depth:** $d = \text{near} + \text{sigmoid}(x) \cdot (\text{far} - \text{near})$, with $\text{near} = 0.2$ and $\text{far} = 400$, ensuring depth values lie within a physically plausible range.
- **Quaternion:** No activation is applied.

The motion head predicts 12-dimensional third-order motion properties. For each order $l \in \{1, 2, 3\}$, it outputs a scalar velocity magnitude and a 3-dimensional directional vector, resulting in $4 \times 3 = 12$ dimensions.

The semantic head produces a 64-dimensional semantic feature map intended for novel-view feature rendering. This feature map is subsequently refined by an auxiliary MLP decoder that expands its dimensionality from 64 to 512, ensuring compatibility with the LSeg feature space.

2. Why We Choose LSeg

Within the framework of feature distillation, the capability of teacher features is crucial for the final semantic reconstruction performance of the model. In order to enable our features to possess language alignment capability, we select three types of CLIP-related features and conduct experimental comparisons.

□ **MaskCLIP** [2]: Standard CLIP [6] computes similarity only between text and global visual feature during contrastive learning, resulting in the local visual features not being strictly aligned with text. MaskCLIP is a CLIP variant that achieves alignment between local features and text. However, its features undergo significant spatial downsampling, yielding semantically condensed but geometrically distorted representations. In SAB3R [1], MaskCLIP features are processed via FeatUp [3], which upsamples the low-resolution features to restore geometric fidelity. As shown in the second row of Figure 1, in our experiments, this upsampling method can accurately restore the edges of some instances. However, it tends to cause feature confusion, which easily interferes with feature learning.

□ **SAM-CLIP**: We use SAM-CLIP to denote instance-level CLIP features obtained by extracting CLIP features from SAM-segmented regions. Similar to PE3R [4], we use SAM to segment images and extract CLIP features from the segmented regions. Meanwhile, we employ SAM2 [7] to perform instance ID alignment across different frames and views, and conduct feature aggregation for identical instances. As shown in the third row of Figure 1, benefiting from the segmentation prior, the instance boundaries in the SAM-CLIP feature map are extremely clear. However, the segmentation prior also introduces several drawbacks, such as the presence of empty feature regions (where no instances are segmented) and potential feature jumps across frames (resulting from jumps in instance segmentation results across frames). These drawbacks make it less suitable for 4D reconstruction that involves a temporal dimension. Additionally, the complex processing procedure leads to very low efficiency in extracting such features.

□ **LSeg-CLIP** [5]: LSeg-CLIP extends CLIP to the semantic segmentation task and enables text query-driven semantic segmentation. Specifically, LSeg-CLIP calculates the similarity between visual features and text features of various categories, which serves as the basis for category classification. Additionally, LSeg-CLIP adopts a visual encoder with low-magnification downsampling, allowing its features to retain a certain degree of geometric structure. As

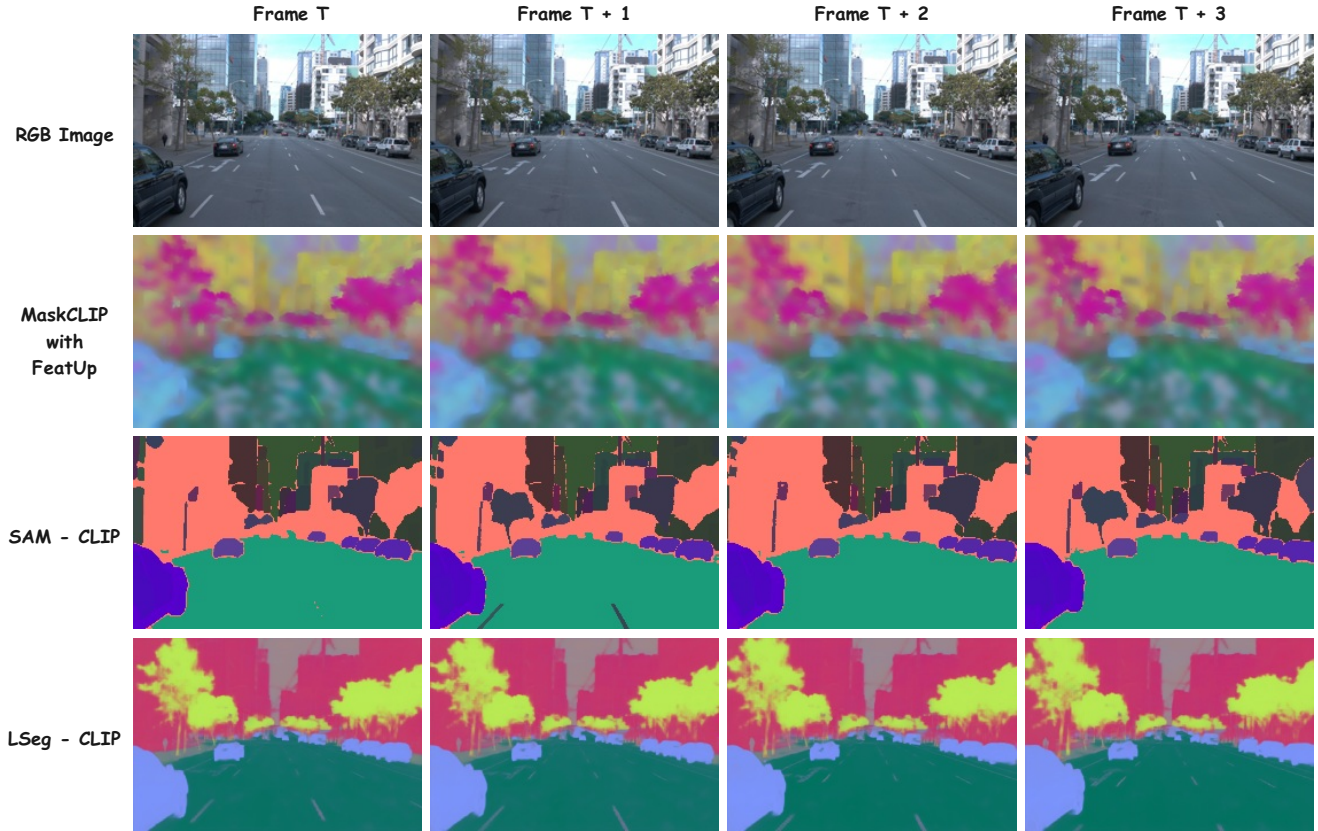


Figure 1. Comparison of different language-aligned features. The first row is the input consists of a set of images from adjacent frames, the next three rows correspond to the three types of features for each frame, respectively.

shown in the last row of Figure 1, in our experiments, LSeg-CLIP features exhibit strong semantic expression capabilities, with sufficiently unified semantics for each category. Although its instance boundaries are less clear than those of SAM-CLIP, it possesses advantages that SAM-CLIP lacks: the absence of feature-less regions, the continuity between inter-frame features, and the efficiency of feature extraction. After comprehensive comparison, we adopt LSeg-CLIP features as our teacher features to endow Gaussian primitives with the capability of semantic reconstruction.

3. More Experiment Results

We present additional image and video results captured from novel viewpoints across a diverse set of dynamic scenarios. These include relatively simple cases—such as scenes with sparse moving objects exhibiting smooth motion (see Figure 2)—as well as highly complex environments characterized by multiple simultaneously moving people and heterogeneous dynamic objects (see Figures 3–5). Further supplementary results, provided in the accompanying folder, corroborate that our method, SLARM, consistently delivers robust and high-fidelity reconstructions

across this spectrum of scene complexity. Notably, SLARM preserves strong temporal coherence, accurate geometry, and photorealistic detail under both subtle motions and intricate multi-agent interactions, underscoring its generalizability and practical efficacy in real-world dynamic settings.

Additional experimental results are available in the SLARM-web directory; opening `index.html` in a web browser provides access to the videos and images.

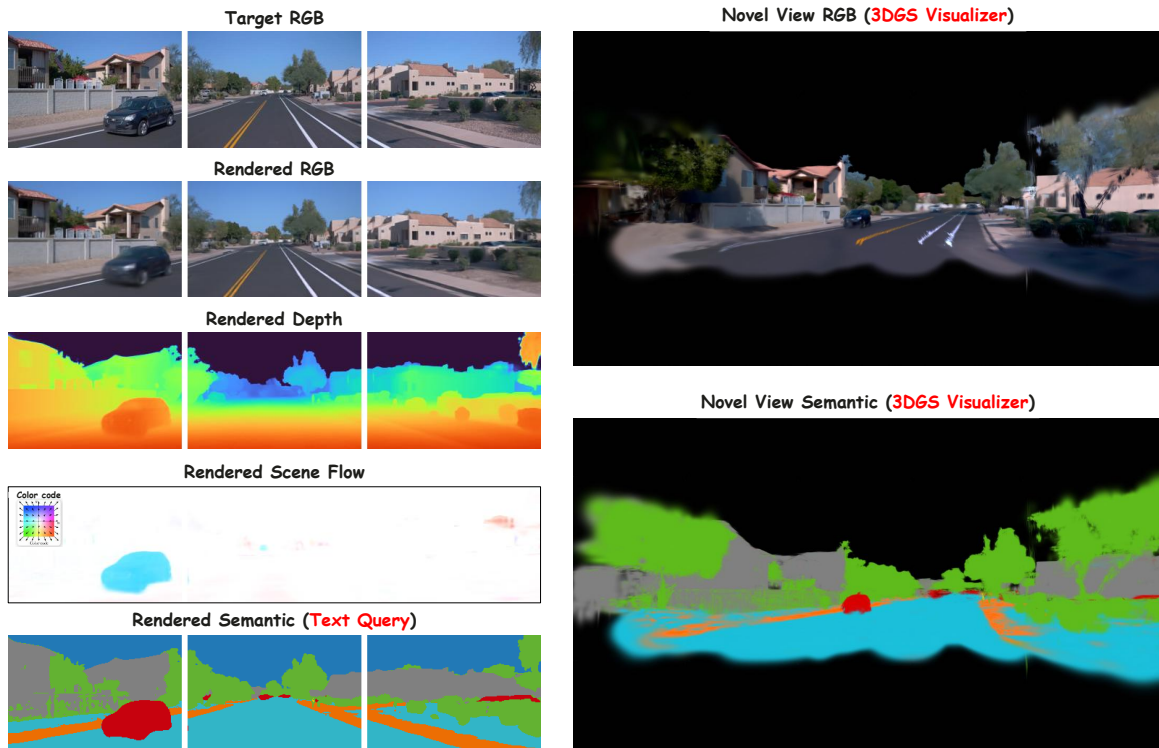


Figure 2. Qualitative results on a simple outdoor scene: left shows rendered RGB, depth, 3D scene flow, and semantic map from predicted 4DGS; right displays a novel view of the 4DGS in a 3DGS visualizer.

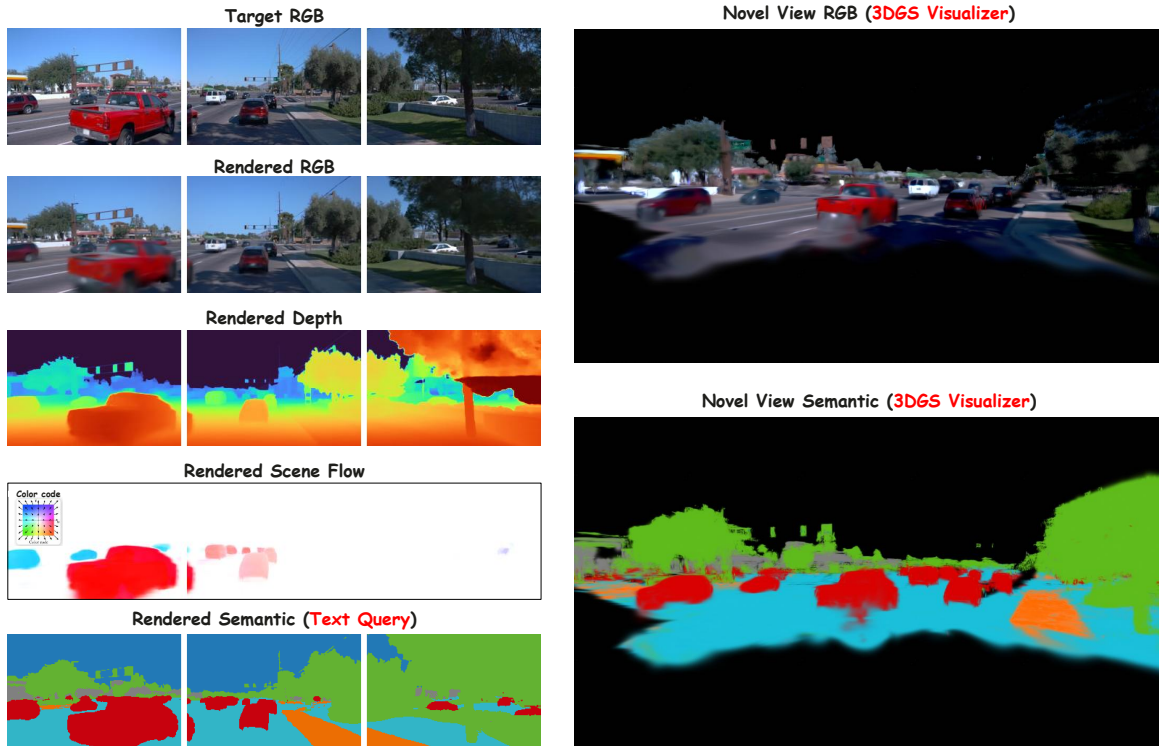


Figure 3. Qualitative results on a complex outdoor scene: left shows rendered RGB, depth, 3D scene flow, and semantic map from predicted 4DGS; right displays a novel view of the 4DGS in a 3DGS visualizer.

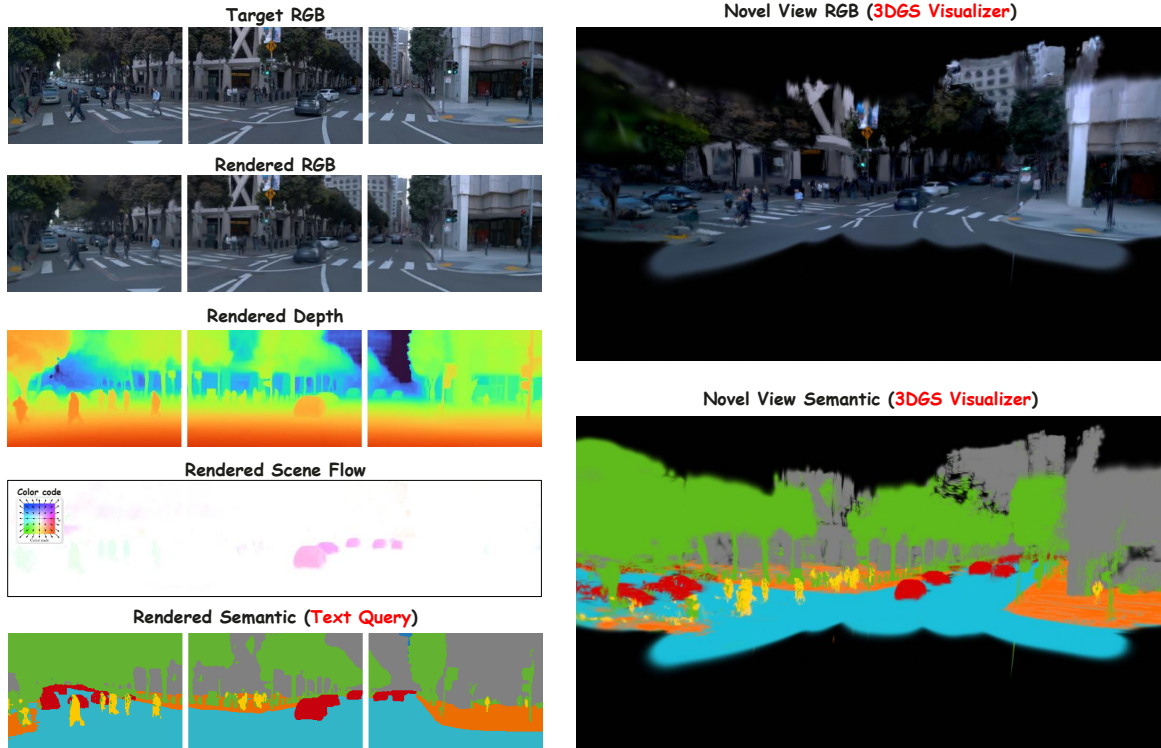


Figure 4. Qualitative results on a complex outdoor scene: left shows rendered RGB, depth, 3D scene flow, and semantic map from predicted 4DGS; right displays a novel view of the 4DGS in a 3DGS visualizer.

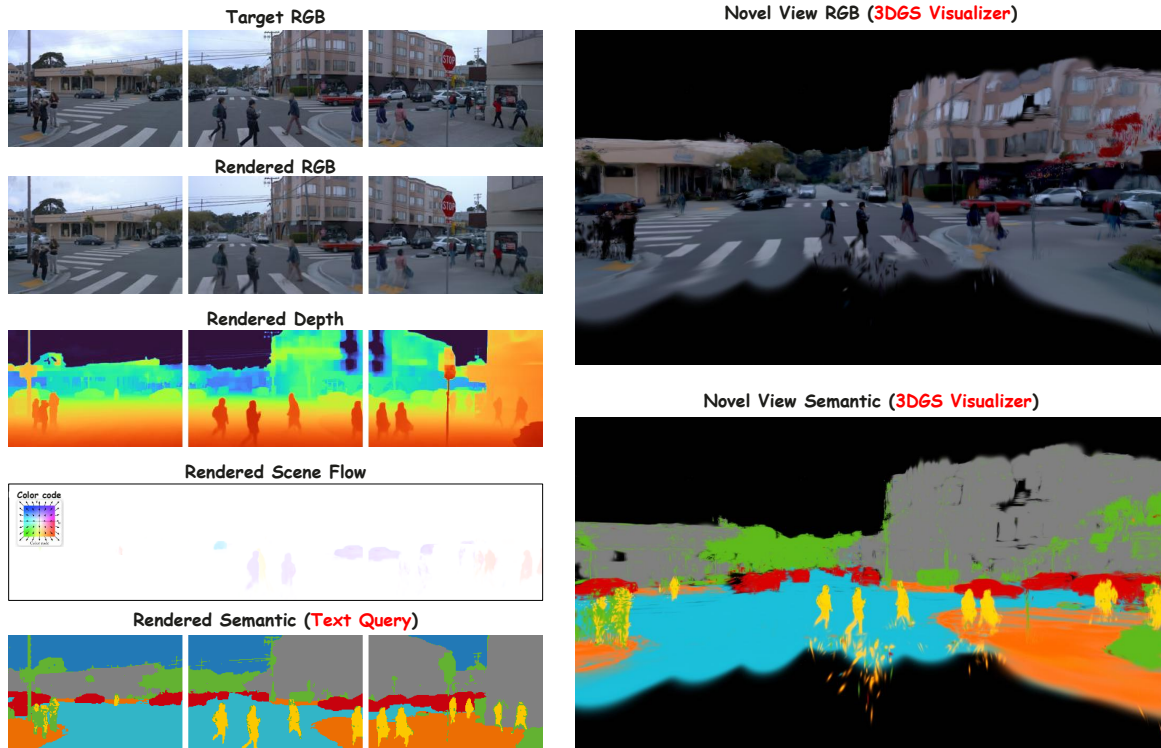


Figure 5. Qualitative results on a complex outdoor scene: left shows rendered RGB, depth, 3D scene flow, and semantic map from predicted 4DGS; right displays a novel view of the 4DGS in a 3DGS visualizer.

References

- [1] Xuweiyi Chen, Tian Xia, Sihan Xu, Jianing Yang, Joyce Chai, and Zezhou Cheng. Sab3r: Semantic-augmented backbone in 3d reconstruction. *arXiv preprint arXiv:2506.02112*, 2025. [1](#)
- [2] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10995–11005, 2023. [1](#)
- [3] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024. [1](#)
- [4] Jie Hu, Shizun Wang, and Xinchao Wang. Pe3r: Perception-efficient 3d reconstruction. *arXiv preprint arXiv:2503.07507*, 2025. [1](#)
- [5] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. [1](#)
- [6] 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. [1](#)
- [7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. [1](#)
- [8] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. [1](#)
- [9] 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. [1](#)