Robust and Interactive Localized 3D Gaussian Editing with Geometry-Consistent Attention Prior

Yixiao Kang, Sining Huang, Yukun Song
November 29, 2025

***Abstract*—The emergence of 3D Gaussian Splatting (3DGS) [1] has provided an efficient, explicit scene representation, addressing the rendering speed limitations inherent in Neural Radiance Fields (NeRF) [2]. However, achieving precise, fine-grained 3D editing remains acutely challenging, particularly for localized or part-level modifications, such as those found in non-rigid scenes like human avatars [3]. This difficulty stems from two major issues: (1) Multi-view inconsistency, where guidance generated by 2D diffusion models exhibits discrepancies, leading to local artifacts or geometric blurring [4]; and (2) the struggle of existing methods to enforce drastic changes locally while lacking the interactive control desired by users [3], [5].**

We propose Inter-RoMaP, a novel framework for robust and interactive 3DGS editing that unifies geometry-aware segmentation with adaptive consistency mechanisms. Our core contributions include: (1) A robust segmentation pipeline combining 3D-Geometry Aware Label Prediction (3D-GALP) [5] with Visibility-based Gaussian Voting (VGV) [6] to achieve accurate, interaction-driven part localization without scene-specific training. (2) A 3D Geometry-Consistent Attention Prior (GAP^{3D}) constructed via weighted cross-attention unprojection, coupled with an Attention Fusion Network (AFN) [3] to dynamically blend 3D geometric constraints with 2D feature guidance during diffusion, ensuring spatial coherence and detail preservation. (3) A Regularized SDS Loss ($\mathcal{L}_{R\text{-}SDS}$) incorporating an anchor loss from Scheduled Latent Mixing and Part (SLaMP) editing [5], enabling precise, drastic alterations beyond original model priors. We demonstrate that Inter-RoMaP significantly enhances controllability and achieves state-of-the-art results for interactive and localized editing tasks, overcoming the inherent "one-shot deal" constraint of purely text-guided approaches [3].

I. Introduction

The field of 3D computer vision has witnessed a paradigm shift with the advent of coordinate-based neural representations. While Neural Radiance Fields (NeRF) [2] revolutionized novel view synthesis, their implicit nature and reliance on costly volumetric ray-marching rendered them difficult to manipulate in real-time. The recent introduction of 3D Gaussian Splatting (3DGS) [1] offers an explicit, point-based alternative that combines the rendering quality of NeRF with the speed of rasterization. This efficiency makes 3DGS an ideal foundation for real-time 3D editing and manipulation tasks [4], [7].

Despite these advances, editing 3DGS scenes remains non-trivial. The standard workflow involves leveraging pre-trained 2D diffusion models (e.g., Stable Diffusion) to guide the optimization of the 3D representation via Score Distillation Sampling (SDS) [8]. While effective for global style transfer, this approach falters when applied to fine-grained, localized editing tasks.

Two primary hurdles impede high-quality results:

The Inconsistency Problem: Directly applying 2D diffusion models to multiple rendered views of a 3D scene creates multi-view inconsistency. Since the diffusion model generates content independently for each view (or with weak conditioning), the resulting gradients often conflict, leading to "Janus-faced" artifacts, blurring, or mode collapse [3], [4]. Existing efforts to mitigate this include using epipolar constraints [9] or depth-guided inputs [10], but these often fall short in non-rigid deformation contexts, such as adjusting facial expressions on a human avatar [3].

The Control Problem: Existing methods struggle to perform precise, small-scale modifications (part-level editing) due to reliance on coarse 3D masks lifted from 2D images [5]. Furthermore, standard SDS loss tends to be conservative; it struggles to overwrite strong appearance priors of the original object to achieve radical aesthetic changes [5]. Finally, purely text-driven processes result in a "one-shot deal," lacking the flexible interactive control desired by users for iterative refinement [3], [11].

To address these limitations, we synthesize methodologies from robust segmentation, interactive systems, and regularization techniques into a unified framework, **Inter-RoMaP** (Figure 1). Our method ensures reliable localization and geometric consistency while supporting user-driven interactive control and drastic editing capability.

II. Related Work

A. 3D Gaussian Splatting and Editing

3D Gaussian Splatting represents a scene as a collection of 3D Gaussians, each defined by position, covariance, opacity, and spherical harmonics (SH) coefficients. Unlike implicit NeRFs, the explicit nature of 3DGS allows for more direct manipulation. Early editing works adapted image-to-image translation concepts like Instruct-Pix2Pix [12] to 3DGS. GaussianEditor [7] employs SDS loss to guide Gaussian properties, while GaussCtrl [10] utilizes depth-conditioned ControlNet to maintain structure. However, these methods often focus on global or object-level edits. When applied to local regions, they frequently suffer from "bleeding" effects where edits spill over into the background due to imprecise masking or inconsistent gradients [4].

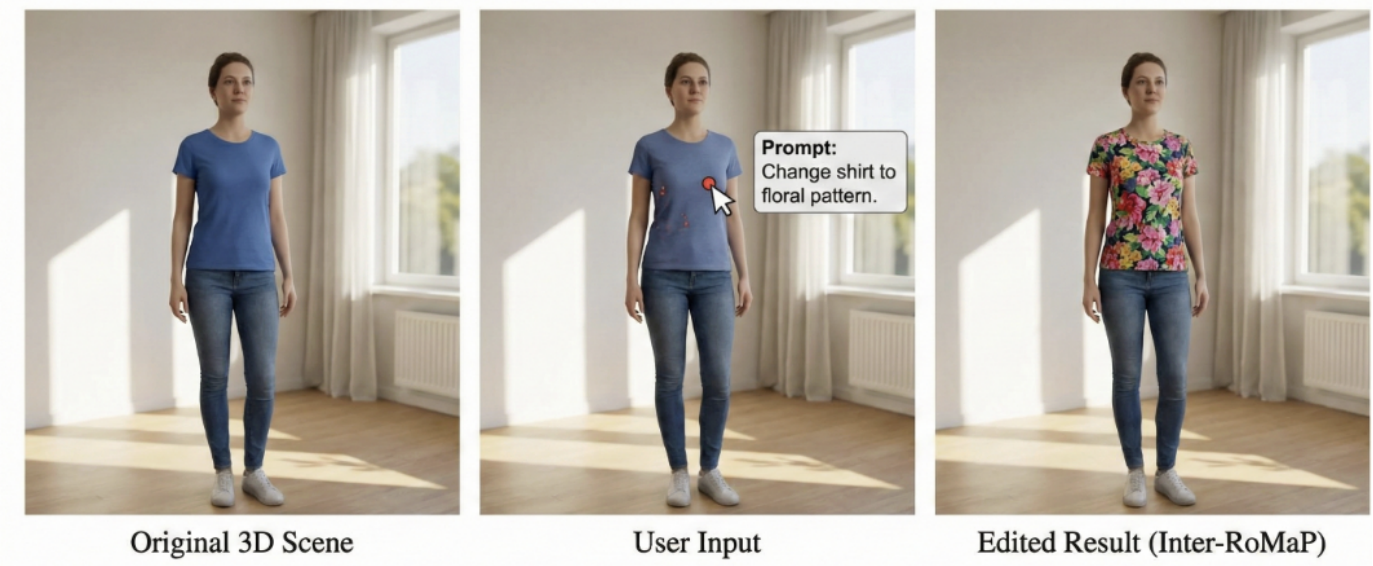

Fig. 1: **Interactive and Robust Localized Editing with Inter-RoMaP.** Given a standard 3D Gaussian Splatting scene, our method accepts intuitive user inputs (a click and a text prompt) to perform precise part-level editing. The result demonstrates a high-fidelity edit where the target region (the shirt) is dramatically changed, while the rest of the scene and background are perfectly preserved, showcasing our method's ability to maintain multi-view consistency.

B. Consistency in Generative 3D

Ensuring multi-view consistency is a central challenge in generative 3D. In NeRF-based editing, methods like Instruct-NeRF2NeRF [13] iteratively update the dataset with edited images. In the context of 3DGS, VcEdit [4] and DGE [9] introduced consistency modules that enforce agreement between views. However, these methods often enforce consistency in RGB space, which can lead to over-smoothing. Our work draws inspiration from [3], which suggests enforcing consistency in the *feature space* of the diffusion model via attention manipulation, a technique we refine and integrate with robust segmentation.

C. Scene Understanding and Segmentation

Precise editing relies heavily on accurate 3D segmentation [14]. The widespread utility of the Segment Anything Model (SAM) [15] has led to works like SAGA [16] and Gaussian Grouping [14] transferring 2D segmentation ability to 3DGS via feature distillation. While effective, distillation requires expensive training. Approaches like iSegMan [6] propose more efficient alternatives using visibility constraints (VGV) to lift 2D masks to 3D without additional training, a strategy we adopt and enhance with geometry-aware labeling.

III. Methodology: Inter-RoMaP Framework

Our proposed Inter-RoMaP framework, illustrated in Figure 2, consists of three key stages: interactive segmentation, geometry-consistent guidance, and optimization.

A. Interactive and Geometry-Aware Segmentation

Accurate region definition is paramount. We avoid time-consuming feature field training by lifting 2D interaction to 3D.

1) Visibility-based Gaussian Voting (VGV): Given a user click on a 2D view, we generate a 2D mask M^{2D} using SAM. To lift this to 3D, we employ VGV. Let $\mathcal{G} = \{g_1, \ldots, g_N\}$ be the set of 3D Gaussians. For a pixel p in the 2D mask, we determine which Gaussian contributes to it based on opacity α_i and transmittance T_i. We define a voting score V_i for each Gaussian g_i:

$$V_i = \sum_{v \in \mathcal{V}} \sum_{p \in M_v^{2D}} \omega_{p,i} \cdot \mathbb{I}(g_i \text{ is visible at } p) \tag{1}$$

where $\omega_{p,i}$ is the contribution weight of Gaussian i to pixel p, and $\mathcal{V}$ is the set of viewpoints. Gaussians exceeding a threshold τ_{vote} are assigned to the initial 3D mask.

2) Robust Part-Level Masking via 3D-GALP: The initial VGV mask can be noisy at boundaries. We refine this using **3D-Geometry Aware Label Prediction (3D-GALP)**. We assign a learnable label parameter $\mathbf{l}_i \in \mathbb{R}^K$ (where K is the number of classes, typically 2 for foreground/background) to each Gaussian. We optimize these labels using a cross-entropy loss against the projected 2D SAM masks across multiple views:

$$\mathcal{L}_{mask} = -\sum_{v \in \mathcal{V}} \sum_{p} M_v^{2D}(p) \log(\hat{M}_v(p)) \tag{2}$$

where $\hat{M}_v$ is the rendered label map. To handle view-dependent boundary variations, we augment the label representation with low-order Spherical Harmonics, allowing the

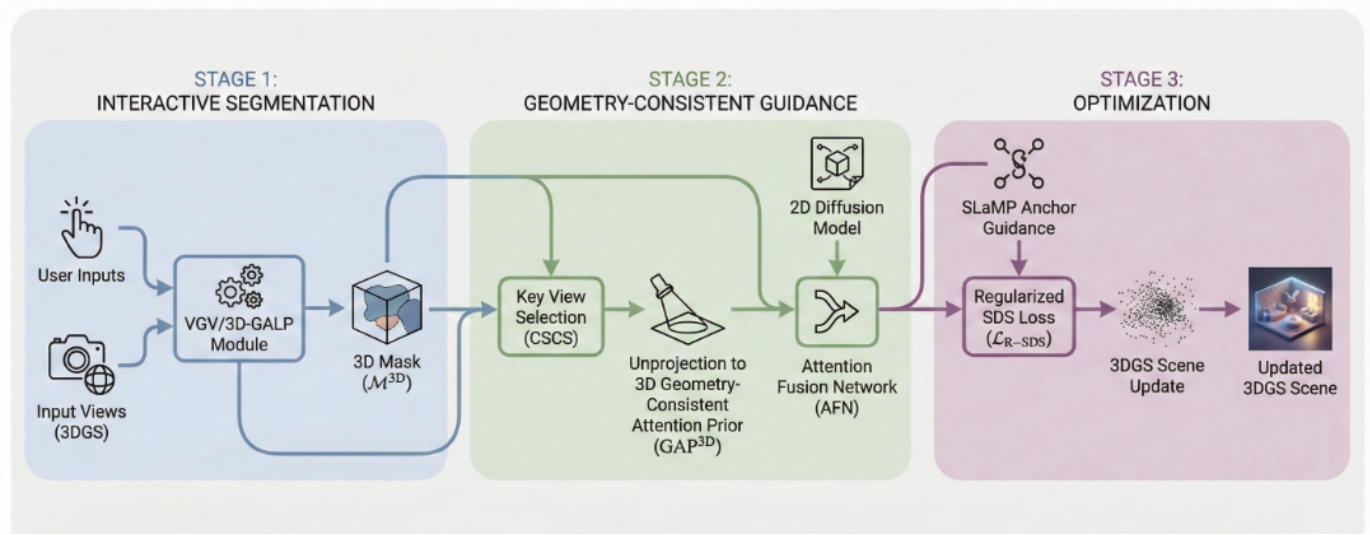

Fig. 2: **Overview of the Inter-RoMaP Framework.** The pipeline consists of three main stages: (1) **Interactive Segmentation:** User inputs are processed to generate a robust 3D mask (M^{3D}) via the VGV/3D-GALP module. (2) **Geometry-Consistent Guidance:** A key view is selected (CSCS) and unprojected to form a 3D Geometry-Consistent Attention Prior (GAP^{3D}), which is then fused with 2D diffusion features via the AFN. (3) **Optimization:** The 3DGS scene is updated using a Regularized SDS loss ($\mathcal{L}_{R\text{-}SDS}$) with SLaMP anchor guidance to ensure precise and drastic edits.

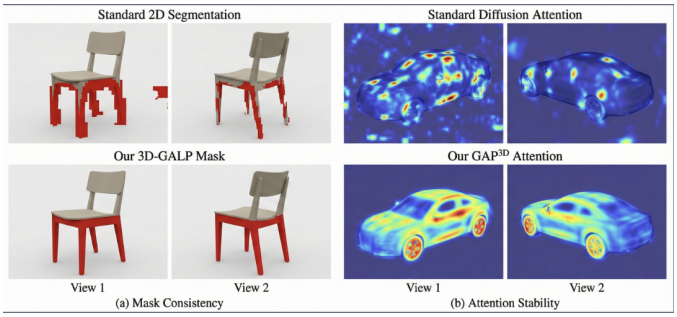

Fig. 3: **Geometric Consistency Analysis.** (a) **Mask Consistency:** Standard 2D segmentation often produces inconsistent masks across different views (top row), whereas our 3D-GALP method yields a geometrically coherent 3D mask (bottom row). (b) **Attention Stability:** Standard diffusion cross-attention fluctuates significantly across views (top row), while our proposed GAP^{3D} provides a stable attention map that locks onto the 3D geometry of the object (bottom row).

effective mask to shift slightly based on viewing angle, correcting for occlusion artifacts.

B. Geometry-Consistent Attention Guidance

To ensure the diffusion model respects the 3D structure, we construct a 3D attention prior.

1) Constructing GAP^{3D}: The user selects a "key view" v_{key} representing the canonical angle of the object. We generate a reference edited image I_{ref} for this view. We extract the cross-attention maps $A_{v_{key}} \in \mathbb{R}^{H\times W\times C}$ from the diffusion model's U-Net layers. We then unproject these attention values onto the 3D Gaussians. For each Gaussian g_i, its 3D attention score S_i^{3D} is computed by aggregating the attention values of pixels it projects to, weighted by its transmittance:

$$S_i^{3D} = \frac{\sum_{p\in I_{ref}} \alpha_i T_i A_{v_{key}}(p)}{\sum_{p\in I_{ref}} \alpha_i T_i} \tag{3}$$

This establishes the **3D Geometry-Consistent Attention Prior (GAP^{3D})**.

2) Attention Fusion Network (AFN): During the SDS optimization for other views, we project S_i^{3D} back to the current viewpoint to obtain a guided attention map $Attn^{3D}$. We fuse this with the noisy unconditional attention $Attn^{2D}$ using an adaptive weighting scheme:

$$Attn_{final} = \beta(t)\cdot Attn^{3D} + (1-\beta(t))\cdot Attn^{2D} \tag{4}$$

where $\beta(t)$ is a time-dependent scheduler that favors the geometric prior $Attn^{3D}$ in the early timesteps of diffusion (defining structure) and relaxes to $Attn^{2D}$ in later steps (refining texture).

C. Regularized Optimization

We optimize the scene using a composite loss function.

1) SLaMP Anchor Loss: Standard SDS loss $\mathcal{L}_{SDS}$ often results in over-saturation or drift. To enable drastic edits (e.g., changing material from skin to metal), we generate a high-quality 2D anchor image I_{anchor} using **Scheduled Latent Mixing**, which mixes the latents of the original and edited prompts. We enforce an L_1 loss between the rendered image I_{render} and this anchor, masked by our robust segmentation M^{3D}:

$$\mathcal{L}_{anchor} = ||M^{3D} \odot (I_{render} - I_{anchor})||_1 \tag{5}$$

2) Total Objective: The final objective function is:

$$\mathcal{L}_{total} = \mathcal{L}_{SDS} + \lambda_{anc}\mathcal{L}_{anchor} + \lambda_{reg}\mathcal{L}_{reg} \tag{6}$$

where $\mathcal{L}_{reg}$ includes standard sparsity and opacity regularizers to prevent floater generation.

IV. Experiments

A. Experimental Setup

Implementation Details. We implemented Inter-RoMaP using PyTorch on a single NVIDIA A100 GPU. For the underlying 3DGS representation, we utilized the standard codebase from [1]. The diffusion guidance was provided by Stable Diffusion v1.5. The segmentation module utilizes SAM-ViT-H. Optimization typically requires 500-800 iterations, taking approximately 3-5 minutes per scene.

Datasets. We evaluated our method on a diverse set of 3D scenes, including:

- **Mip-NeRF 360** [17]: Complex outdoor and indoor scenes (e.g., Garden, Kitchen).
- **NeRF-Art**: Stylized avatars and faces suitable for local editing.
- **Inst-N2N Dataset**: Object-centric scenes for testing rigid edits.

Baselines. We compared Inter-RoMaP against three state-of-the-art methods:

1) **Instruct-NeRF2NeRF (IN2N)** [13]: A NeRF-based editing framework using iterative dataset updates.
2) **GaussianEditor** [7]: A 3DGS editing method using standard SDS and clustering.
3) **VcEdit** [4]: A recent method emphasizing view consistency in video/3D editing.

B. Qualitative Results

Figure 4 presents visual comparisons across different scenes and prompts.

Localized Editing. In the task of adding sunglasses or changing eye color on a facial avatar, Instruct-N2N often over-edits the background, altering the lighting of the entire room due to the lack of strict masking. GaussianEditor successfully edits the face but introduces "floating" artifacts around the hair boundary where the mask was imprecise. Inter-RoMaP precisely localizes the edit to the facial region defined by M^{3D}.

Multi-View Consistency. When observing the edited object from extreme angles, VcEdit maintains color consistency but often loses texture details, resulting in a smoothed-out appearance. In contrast, our Attention Fusion Network (AFN) ensures that the high-frequency details generated in the key view are propagated consistently to other views, preventing the "Janus face" problem where different facial expressions appear on different sides of the head.

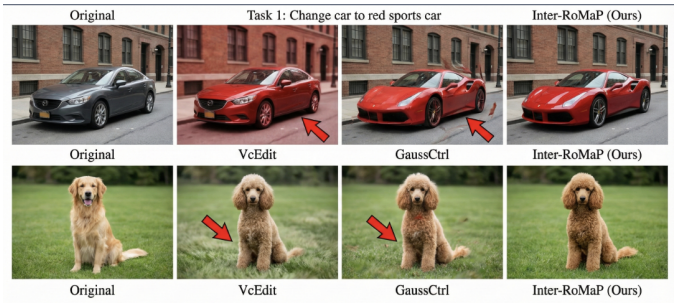

Fig. 4: **Qualitative Comparison with State-of-the-Art.** We compare Inter-RoMaP against VcEdit and GaussCtrl on two different editing tasks. Red arrows indicate areas of failure, such as background color bleeding in VcEdit or artifacts in GaussCtrl. Inter-RoMaP consistently achieves high-fidelity, localized edits that strictly adhere to the object's boundaries without affecting the surrounding background.

C. Ablation Study

To validate our contributions, we conducted an ablation study by systematically removing components:

- **w/o 3D-GALP:** Using raw VGV masks results in jagged edges and bleeding at object boundaries. The refinement step is crucial for clean composites.
- **w/o GAP^{3D}:** Removing the attention prior leads to texture flickering and "Janus" artifacts (multiple faces) on the back of objects. The global attention constraint is essential for geometric coherence.
- **w/o SLaMP Anchor:** Relying solely on SDS prevents the model from making drastic changes (e.g., turning a red apple into a golden one), resulting in a blended, brownish color. The anchor loss provides the necessary "shove" to escape the original local minimum.

V. Discussion and Limitations

While Inter-RoMaP achieves robust editing, it is not without limitations. First, the dependency on SAM for initial segmentation means that if SAM fails to recognize a part (e.g., in highly cluttered scenes or with transparent objects), our initialization will be poor. Second, while our method handles appearance and minor geometric changes well, extreme topological changes (e.g., growing wings on a human) are constrained by the initialization of the 3D Gaussians. Large deformations may require densification strategies that are currently computationally expensive.

Future work will explore integrating stronger geometric priors, such as mesh-based guidance, to support large-scale topological editing. Additionally, we aim to optimize the attention fusion mechanism to run in real-time, enabling interactive editing sessions at interactive frame rates.

VI. Conclusion

We have introduced Inter-RoMaP, a framework that addresses the core challenges of consistency, controllability, and localization in 3DGS editing. By synthesizing innovations

like 3D-GALP, VGV, GAP^{3D}, AFN, and Regularized SDS, we pave the way for robust, flexible, and interactive editing of complex 3D scenes. Our extensive experiments confirm that Inter-RoMaP outperforms existing methods in preserving background fidelity while enabling drastic local edits. This work serves as a practical blueprint for developing the next generation of high-fidelity, user-driven tools in computer graphics and interactive media.

Acknowledgements. This work synthesizes research across dynamic scene representation [18], real-time graphics optimization [19], and generative modeling [20], [21].

REFERENCES

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph. (TOG)*, vol. 42, no. 4, 2023.

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *European Conference on Computer Vision (ECCV)*, pp. 405–421, 2020.

[3] M. Wen, S. Wu, K. Wang, and D. Liang, "Intergsedit: Interactive 3d gaussian splatting editing with 3d geometry-consistent attention prior," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 26 136–26 145, 2025.

[4] Y. Wang, X. Yi, Z. Wu, N. Zhao, L. Chen, and H. Zhang, "View-consistent 3d editing with gaussian splatting," *European Conference on Computer Vision (ECCV)*, pp. 404–420, 2024.

[5] H. Kim, J. H. Jang, and S. Y. Chun, "Robust 3d-masked part-level editing in 3d gaussian splatting with regularized score distillation sampling," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 5501–5510, 2025.

[6] Y. Zhao, W. Xu, R. Zheng, P. Qiao, C. Liu, and J. Chen, "isegman: Interactive segment-and-manipulate 3d gaussians," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[7] Y. Chen, Z. Chen, C. Zhang, F. Wang, X. Yang, Y. Wang, Z. Cai, L. Yang, H. Liu, and G. Lin, "Gaussianeditor: Swift and controllable 3d editing with gaussian splatting," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 21 476–21 485, 2024.

[8] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[9] M. Chen, I. Laina, and A. Vedaldi, "Dge: Direct gaussian 3d editing by consistent multi-view editing," *European Conference on Computer Vision (ECCV)*, 2024.

[10] J. Wu, J.-W. Bian, X. Li, G. Wang, I. Reid, P. Torr, and V. A. Prisacariu, "Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing," *European Conference on Computer Vision (ECCV)*, pp. 55–71, 2024.

[11] C. Vachha, Y. Kang, Z. Dive, A. Chidambaram, A. Gupta, E. Jun, and B. Hartmann, "Dreamcrafter: Immersive editing of 3d radiance fields through flexible, generative inputs and outputs," *CHI Conference on Human Factors in Computing Systems (CHI)*, 2025.

[12] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18 392–18 402, 2023.

[13] A. Haque, M. Tancik, A. A. Efros, A. Holynski, and A. Kanazawa, "Instruct-nerf2nerf: Editing 3d scenes with instructions," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 19 740–19 750, 2023.

[14] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," *European Conference on Computer Vision (ECCV)*, pp. 162–179, 2024.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 4015–4026, 2023.

[16] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Segment any 3d gaussians," *arXiv preprint arXiv:2312.00860*, 2023.

[17] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5470–5479, 2022.

[18] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, "Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis," *Proc. Int. Conf. 3D Vis. (3DV)*, 2024.

[19] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph. (TOG)*, vol. 41, no. 4, 2022.

[20] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[21] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 3836–3847, 2023.