EditAnyScene: Text-Driven 3D Scene Local Editing with Gaussian Splatting

Jiawei Zhou[1,*], Yang Liu[1,*], Yufeng Wang[1], Yanning Zhou[2,†], Haoqian Wang[1,†]
[1]Tsinghua University [2]Tencent
zhoujw22@mails.tsinghua.edu.cn, liu-yang22@mails.tsinghua.edu.cn, yufeng-w22@mails.tsinghua.edu.cn
amandayzhou@tencent.com, wanghaoqian@tsinghua.edu.cn

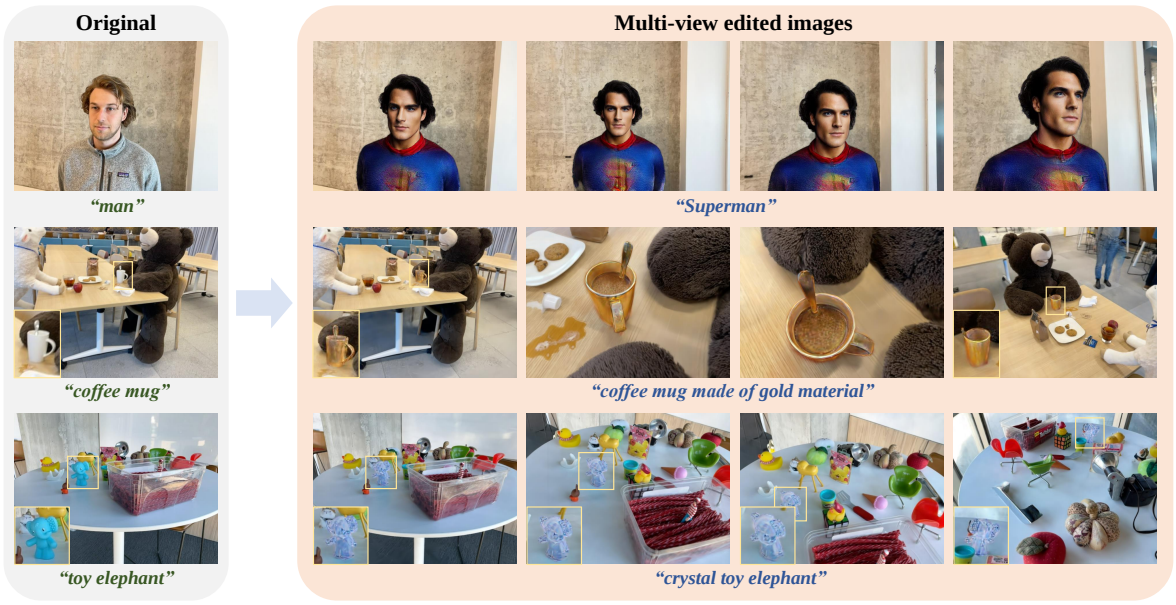

Figure 1. **Multi-view Edited Results.** We present our edited result over a single-object scene "face" and two multi-object scenes "teatime" and "figurines". EditAnyScene can edit objects of any size in both single-object and multi-object scenes, and provide high-quality edited results from any perspective.

Abstract

We present EditAnyScene, a generalizable framework for text-driven local editing of 3D Gaussian Splatting scenes. While existing approaches yield promising results in simple environments, they often deteriorate in complex real-world scenes due to imprecise target localization, limited viewpoint coverage, and inconsistent multi-view editing. EditAnyScene overcomes these challenges through three complementary innovations. First, we construct a 3D Language Field that bridges natural language and spatial understanding, enabling accurate target localization and extraction in cluttered scenes. Further, we develop an object-centric optimal viewpoint sampling strategy that generates dense orbiting trajectories around the identified target, transcending the limitations of initial camera perspectives through comprehensive spatial representation of the object. Building upon this enhanced visual representation, we introduce a multi-view joint editing scheme featuring Global-Local Attention Latent Alignment and Object-Level Attention modules. This coordinated approach enforces consistent geometric and stylistic guidance in latent space while effectively suppressing background interference during foreground editing, ensuring coherent modifications. Extensive experiments demonstrate that EditAnyScene achieves superior editing quality and generalization across diverse scenes, successfully handling scenarios where previous methods fail. Our work represents a promising step toward robust, generalizable 3D scene editing guided by natural language instructions.

* Equal contribution. †Corresponding authors.

1. Introduction

Recent advancements in 3D representations, particularly Neural Radiance Fields (NeRF)[18] and 3D Gaussian Splatting (3DGS)[8], have significantly advanced real-world scene reconstruction. Building on these innovations, text-driven 3D scene editing [3, 4, 6, 7, 11, 13, 15, 17, 20, 21, 23, 25] has gained attention for applications in virtual reality and gaming. However, editing specific objects in complex 3D environments while preserving scene consistency remains highly challenging.

Instruct NeRF2NeRF[4] utilizes image-conditioned diffusion models to guide NeRF-based editing, achieving multi-view consistency but often leading to unintended background changes due to implicit target localization. GaussianEditor [2] takes advantage of 3DGS's discrete structure for object-specific editing via inverse rendering and 2D mask projections. Meanwhile, GaussCtrl [22] improves consistency through depth-guided geometry alignment and cross-view attention.

Despite these advances, existing methods face significant challenges when scaling to multi-object scenarios: **Inconsistent target localization**: Open-vocabulary 2D segmentation models (e.g., Lang-SAM [10, 12, 14]) frequently fail to maintain multi-view consistency. **Suboptimal viewpoint selection**: Existing approaches often lack mechanisms to center training views on relevant objects, limiting editing precision. **Multi-view consistency issues**: Heavy reliance on fixed reference frames [16, 22] can result in artifacts such as blurring from mismatched perspectives.

To address these challenges, we propose EditAnyScene, a novel framework for precise text-driven editing of 3DGS Scenes. First, a 3D Language Field aligns semantic embeddings with Gaussian features by jointly training spatial and language fields, ensuring consistent 3D localization. Second, we propose an object-centric perspective generation strategy that selects optimal viewpoints based on object center and camera distribution, enhancing editing coverage. Finally, we introduce two attention mechanisms: the Global-Local Attention Latent Alignment Module (GLAM) ensures visual consistency through cross-view attention with global and neighboring frames, while the Object-Level Attention Module (OAM) focuses foreground processing on object regions, reducing background interference.

Experiments show that EditAnyScene significantly outperforms previous methods in editing precision, consistency, and visual quality. The contributions of this work are summarized as follows:

- An object-centric perspective generation strategy for optimal editing viewpoints.
- Novel attention mechanisms (GLAM and OAM) for enhanced multi-view consistency.
- Extensive evaluations demonstrating state-of-the-art performance across diverse 3D editing tasks.

2. Method

We present EditAnyScene, a framework for text-driven 3D editing of 3DGS scenes. Fig.2 illustrates our pipeline, which jointly optimizes language and geometric fields for accurate target localization, generates object-centric perspectives for optimal viewpoint selection, and applies specialized attention mechanisms for multi-view consistent modification.

2.1. Language Gaussian for 3D localization

We extend 3DGS by integrating learnable language features distilled from image CLIP embeddings into each Gaussian. Unlike previous methods [19] that train geometry and language fields sequentially, we propose a joint training strategy where both fields are optimized simultaneously:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}}\left(\boldsymbol{C}, I_{\text{ori}}\right) + \lambda_{\text{lang}}\mathcal{L}_{\text{lang}}(\boldsymbol{F}, \boldsymbol{H}) \tag{1}$$

where $\boldsymbol{C}$, I_{ori}, $\boldsymbol{F}$, $\boldsymbol{H}$ represent rendered RGB image, original image, rendered semantic features, and image CLIP embeddings. λ_{lang} balances RGB reconstruction and language learning.

During the Gaussian growth process, new points inherit language features from parent points, enhancing spatial continuity of the semantic field. Simultaneously, we prevent semantic branch gradients from updating geometric parameters to maintain reconstruction quality. This approach enables accurate target localization by measuring cosine similarity between CLIP embeddings and Gaussian features, generating precise 3D masks even in complex scenes.

2.2. Optimal Editing Perspective Generation

Previous 3D editing methods struggle with complex scenes where target objects are not well-centered, leading to localization errors or insufficient editing. We address this by generating optimal viewpoints specifically focused on the target object. First, we filter training viewpoints based on the angular deviation between the camera-to-object vector and the camera's field of view:

$$V_{train} = \{cam_i | \alpha_x^i < \frac{fov_x}{2} \cdot k_\alpha \, , \alpha_y^i < \frac{fov_y}{2} \cdot k_\alpha\} \tag{2}$$

where α_x^i and α_y^i are projection angles of camera-to-object vector on XZ/YZ planes, fov_x and fov_y are camera field of view angles, and $k_\alpha \in [0, 1]$ controls filtering strictness.

Second, we generate dense spherical viewpoints $V_{spherical}$ around the target. We align the spherical trajectory with the object's coordinate system and calculate the optimal camera radius based on existing viewpoints. For azimuth angles, we employ DBSCAN to exclude outliers and apply kernel density estimation to determine visible azimuth segments. This approach ensures comprehensive coverage of the target object from optimal perspectives.

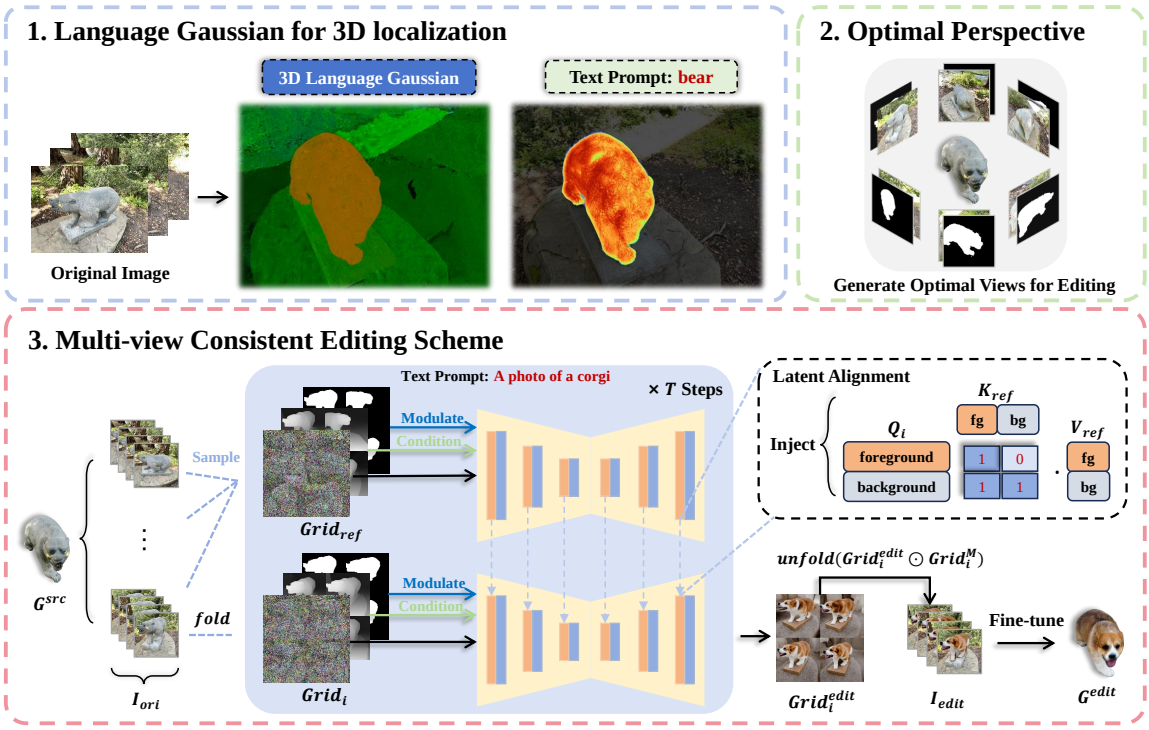

Figure 2. **Overview of EditAnyScene.** First, we reconstruct the Gaussian language field alongside the geometry field from the input images (Sec.2.1). Then, we generate intensive perspective supervision for 2D editing (Sec.2.2). Finally, we use the depth-guided ControlNet [24] in combination with a Global-Local Attention Latent Alignment module (GLAM) and an Object-Level Attention Module (OAM) to achieve multi-view consistency (Sec.2.3).

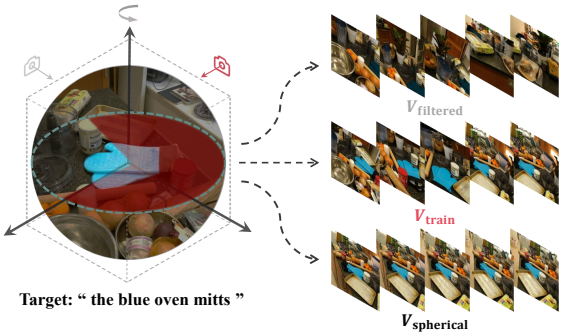

Figure 3. Visualization of Optimal Editing Perspective Generation. Red fan-shaped region indicates the valid azimuth range. $V_{filtered}$ shows filtered-out views with typically poor editing quality, V_{train} displays the selected training views, and $V_{spherical}$ represents the generated continuous spherical viewpoints.

2.3. Multi-view Consistent Editing Scheme

To utilize the continuous and dense characteristics of generated viewpoints, we introduce two complementary modules that actively enhance multi-view consistency throughout the editing process.

Global-Local Attention Latent Alignment Module addresses limitations of previous cross-frame attention methods that are sensitive to reference selection. Our approach organizes sequential images into grid layouts $Grid^{(1,\dots,T)} \in \mathbb{R}^{(nW)\times(nH)\times 3}$, where $T = n \times n$, enabling simultaneous processing that leverages transformer self-attention for temporal connections.

For global consistency, we extract representative frames to construct a reference grid $Grid_{ref}$. During diffusion, we compute cross-frame attention between local grid latents z_i^t and the reference grid, blending it with original self-attention:

$$\text{AttnAlign} = \lambda \cdot \text{Attn}_{i,i} + (1-\lambda) \cdot \text{Attn}_{i,\text{ref}} \quad (3)$$

where $\lambda \in [0,1]$ balances attention types and $\text{Attn}_{i,j}$ represents attention between latents z_i and z_j. This hybrid approach ensures both local geometric accuracy within grids and global stylistic consistency across trajectories, reducing reference view sensitivity.

Object-Level Attention Module prevents background interference during foreground editing by restricting attention operations to within the target object region. Specifically, we first project the previously extracted 3D mask onto various views to create a normalized foreground heatmap H, which is then thresholded to obtain a 2D mask M. The 2D mask is subsequently organized into a grid format. Formally, for the i^{th} local grid, the corresponding mask is ex-

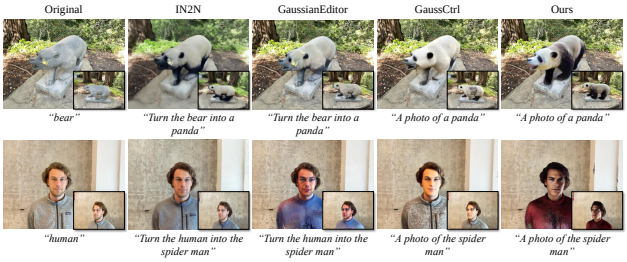

Figure 4. **Qualitative Results on IN2N [4] Dataset.** Compared to baseline methods, EditAnyScene achieves better text prompt alignment and multi-view consistency.

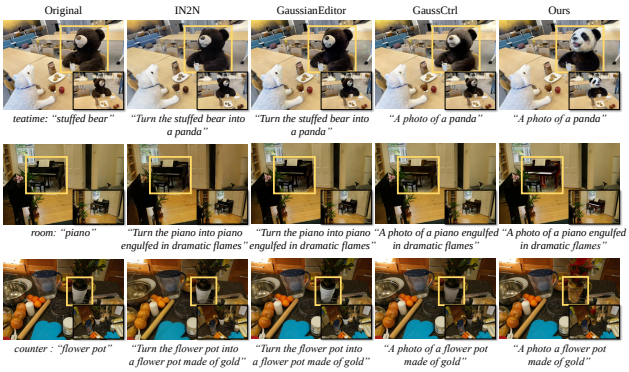

Figure 5. **Qualitative Results on Mip-NeRF360 [1] and LERF [9].** We show the edited results in multi-object scenes. EditAnyScene provides high-quality edited results while baseline methods fail in almost all cases.

pressed as $M_i^{(1,\dots,T)} \in \mathbb{R}^{(nW)\times(nH)\times 1}$. The attention module is then defined as follows:

$$\text{ObjAttn}_{i,j} = \text{Softmax}\left(\frac{Q_i[M_i]\cdot K_j[M_j]^\top}{\sqrt{d}}\right) V_j[M_j] \quad (4)$$

where $Q_i[M_i]$ denotes the selection of all elements in Q_i for which the corresponding values in the mask M_i are equal to 1. The same selection process applies to K_j and V_j.

3. Experiments

3.1. Settings

We compare EditAnyScene with three state-of-the-art techniques: IN2N [5], GaussianEditor [2], and GaussCtrl [22], using $CLIP$ Similarity Score to evaluate visual correspondence between 3D edits and prompts. For our evaluation, we collect diverse scenes from multiple datasets including simple scenes (both 360-degree and forward-facing views from IN2N [4]) and complex scenes from Mip-NeRF360 [1] and LERF [9].

For implementation, we build upon [2], incorporating 3D language Gaussians following LangSplat [19] with $\lambda_{\text{lang}} = 1.0$ for joint training of geometry and language fields. We set $k_\alpha = 0.6$ for optimal view generation. For 2D image editing, we use a grid size of $T = 2 \times 2$ and employ Stable Diffusion v1.5 with depth-conditioned ControlNet (guidance scale 7.5), setting $\lambda_{\text{ref}} = 0.6$ for self-attention latents injection. Our method takes 10-15 minutes per scene on a NVIDIA RTX 3090Ti.

Method	bear	face	teatime	figurines	counter	room	avg
IN2N [4]	0.2545	0.2280	0.2354	0.2406	0.2516	0.2403	0.2420
GaussianEditor [2]	0.2435	0.2426	0.2310	0.2435	0.2482	0.2395	0.2401
GaussCtrl [22]	0.2730	0.2368	0.2294	0.2392	0.2453	0.2364	0.2376
Ours	**0.2752**	**0.2541**	**0.2459**	**0.2550**	**0.2553**	**0.2457**	**0.2505**

Table 1. **Quantitative Evaluation on Simple and Complex Scenes.** We compare the average $CLIP$ Similarity Score of diverse prompts across different scenes. The left section shows simple scenes from the IN2N dataset [4], while the middle section presents results on complex scenes. EditAnyScene consistently performs better than baseline methods in both scenarios.

3.2. Comparative Study

For simple scenes ("bear" and "face" from IN2N[4]), Fig. 4 demonstrates our superior editing quality from different perspectives. Furthermore, we calculate the $CLIP$ Score of diverse prompts on the same scene and show quantitative comparisons in Tab.1. EditAnyScene achieves better text prompt alignment and multi-view consistency in all cases.

For complex multi-object scenes ("teatime" and "figurines" from LERF [9], "counter" and "room" from Mip-NeRF360 [1]), Fig. 5 presents qualitative comparisons. Our method successfully edits target object in these scenes while maintaining high-quality results with multi-view consistency, whereas baseline methods consistently fail to produce satisfactory edits. Due to the difficulty in mapping complex scenes to succinct local descriptive editing prompts, we compute $CLIP$ Score within dilated bounding boxes of projected 3D masks for fair comparison. As shown in Tab. 1, EditAnyScene consistently outperforms baseline methods across all scene types, demonstrating its robust capability in both simple and challenging scenarios.

4. Conclusion

In this paper, we introduce EditAnyScene, an efficient method for 3DGS Local editing. We achieve consistent 3D localization by grounding language features into 3D Gaussians and identifying objects in 3D using text queries. We propose an object-centric perspective generation method to adaptively identify optimal views for editing. Furthermore, we propose a novel multi-view joint editing scheme that effectively facilitates multi-view consistency by employing the Global-Local Attention Latent Alignment Module (GLAM) and the Object-level Attention Module (OAM). In various settings, our method enables more consistent and text-aligned edits than prior works.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 4

[2] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *CVPR*, pages 21476–21485, 2024. 2, 4

[3] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[4] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 2, 4

[5] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 4

[6] Runze He, Shaofei Huang, Xuecheng Nie, Tianrui Hui, Luoqi Liu, Jiao Dai, Jizhong Han, Guanbin Li, and Si Liu. Customize your nerf: Adaptive source driven 3d scene editing via local-global iterative training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6966–6975, 2024. 2

[7] Nazmul Karim, Hasan Iqbal, Umar Khalid, Chen Chen, and Jing Hua. Free-editor: zero-shot text-driven 3d scene editing. In *European Conference on Computer Vision*, pages 436–453. Springer, 2025. 2

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 42(4):1–14, 2023. 2

[9] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *ICCV*, pages 19729–19739, 2023. 4

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2

[11] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *NeurIPS*, 35:23311–23330, 2022. 2

[12] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[13] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5773–5783, 2021. 2

[14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[15] Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024. 2

[16] Chaofan Luo, Donglin Di, Xun Yang, Yongjia Ma, Zhou Xue, Chen Wei, and Yebin Liu. Trame: Trajectory-anchored multi-view editing for text-guided 3d gaussian splatting manipulation. *arXiv preprint arXiv:2407.02034*, 2024. 2

[17] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 2

[18] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[19] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *CVPR*, pages 20051–20060, 2024. 2, 4

[20] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 2

[21] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[22] Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Prisacariu. GaussCtrl: Multi-View Consistent Text-Driven 3D Gaussian Splatting Editing. *ECCV*, 2024. 2, 4

[23] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision*, pages 597–614. Springer, 2022. 2

[24] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 3

[25] Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, and Jiaya Jia. Ref-npr: Reference-based non-photorealistic radiance fields for controllable scene stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4242–4251, 2023. 2