Localized Gaussian Splatting Editing with Contextual Awareness

Hanyuan Xiao[1,2]   Yingshu Chen[3]   Huajian Huang[3]
Haolin Xiong[4]   Jing Yang[1,2]   Pratusha Prasad[1,2]   Yajie Zhao[1,2,*]

[1]University of Southern California   [2]Institute for Creative Technologies
[3]HKUST   [4]University of California, Los Angeles

[1,2]{hxiao, jyang, bprasad, zhao}@ict.usc.edu   [3]{ychengw, hhuangbg}@connect.ust.hk   [4]{xiongh}@ucla.edu
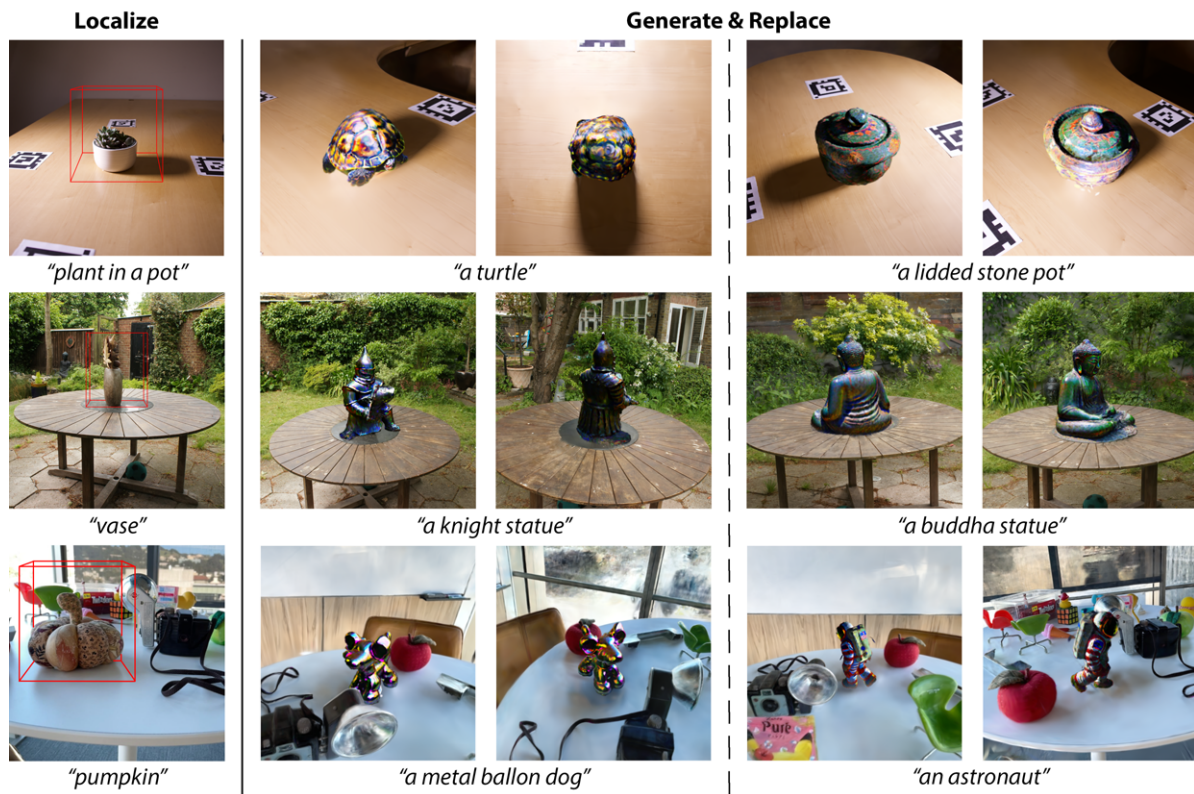
Figure 1. We present a novel pipeline for text-guided localized Gaussian splatting scene editing. It replaces and edits regions of interest based on user text inputs, achieving realistic and high quality visual results that naturally match the original context, including illumination and occlusions. Compared to previous methods, we are the first to enable object replacement with consistent illumination to the surroundings.

Abstract

Recent advancements in text-guided 3D object generation using diffusion priors struggle with illumination inconsistencies when applied to scene editing tasks like object replacement or insertion. To address this, we propose an illumination-aware 3D scene editing pipeline for 3D Gaussian Splatting (3DGS). Our method leverages state-of-the-art 2D diffusion inpainting [56] to handle global illumination context effectively. Specifically, we identify representative anchor views that capture scene-wide illumination, inpaint them using 2D diffusion models, and integrate the results into a coarse-to-fine 3DGS optimization process. In the fine step, we introduce Depth-guided Inpainting Score Distillation Sampling (DI-SDS)

---

*Corresponding Author

to refine geometry and texture details, capitalizing on the diversity of 2D priors. Our approach achieves locally precise edits with globally consistent illumination, demonstrating robustness in real scenes with highlights and shadows. Comparisons show superior results over state-of-the-art text-to-3D editing methods. Project page: https://corneliushsiao.github.io/GSLE.html.

1. Introduction

Recent advancements in text-guided 3D object generation have shown remarkable success by leveraging the power of diffusion models [24, 27, 37, 48, 52, 53]. However, little attention has been paid to text-guided localized scene editing task. Naive insertion of generated 3D objects leads to mismatch of illumination appeared in the scene. Global editing methods that do not include localization module [9, 14, 51] fail by adding color clue to the whole scene or changing undesired regions. Existing solutions for text-guided 3D editing [1, 11, 47, 58] can only achieve minor alternation in geometry or texture of existing objects. Tasks such as context-aware object insertion and object replacement with pronounced change are still underexplored.

In this work, 3D Gaussian Splatting (3DGS) [17] represents both scene and generated object, because 1) it supports efficient optimization in 3D lifting task, and 2) provides explicit representation that enables editing independent of background, in contrast to NeRF representation [33]. To achieve holistically photorealistic 3D editing, it is necessary to take global information such as illumination, occlusions into account. Conventional approaches involving inverse rendering, material generation and physically-based neural rendering are time consuming and unstable [5, 23]. Instead, we observed large-scale diffusion models conditioned on background and geometry information (e.g., depth or camera pose) excel at the inpainting task, particularly in regards to maintaining illumination matching and seamlessly fusing the generated foreground and background with the original image. Building upon this observation, we introduce a 3D scene editing pipeline that considers scene illumination. By advancing algorithm of 3D lifting and score distillation sampling (SDS) [37], our pipeline accomplishes challenging object insertion and replacement by incorporating prior knowledge from both a depth-guided inpainting diffusion model [56] and a 3D-aware diffusion model [27].

For robust and harmonious object synthesis, we introduce Anchor View Proposal (AVP) algorithm, which selects a view that best represents global illumination with the most contrast. This picked view will be inpainted given the input text prompt, working as the initial guidance for the following two-step contextual 3D lifting. By analysis, this inpainted guidance offers heuristic information for illumination-aware object synthesis. Among the

two steps, we first conduct a coarse image-to-3D generation and then an illumination-aware texture enhancement optimization. In the coarse step, we explore score distillation of the 3D-aware diffusion prior from the multi-view diffusion model [27] given the foreground object in the inpainted anchor view. This diffusion prior comprehends essential illumination information from anchor view and guarantees multi-view consistency in 3D object synthesis. In the enhancement step, we propose a new score distillation scheme (DI-SDS) that leverages depth-guided inpainting diffusion prior with background context. This strategy yields multi-view-consistent and contextually illumination-aware object refinement.

To summarize, our main contributions include:

1. We present a comprehensive pipeline to generate objects based on text inputs that seamlessly integrate into a 3D Gaussian Splatting scene. In particular, our method enables objects generation that automatically matches global illumination.

2. We introduce Anchor View Proposal algorithm for automatic selection of a representative view that best represents illumination at a target region within a complex scene. The anchor view selection provides features of scene illumination for harmonious object synthesis.

3. We propose Depth-guided Inpainting Score Distillation Sampling, which embeds both condition of geometry and contextual illumination for object generation and texture enhancement.

2. Related Works

Recent breakthroughs in large-scale vision-language models (VLMs) [22, 40] and text-guided generative models featuring Stable Diffusion [41] have boosted the research field in 2D-to-3D generation and editing, leveraging these powerful priors. We review recent literature on image- and text-guided 3D generation and 3D scene editing, focusing on scenes using radiance field representation.

2.1. 3D Object Generation

Early exploration of text-to-3D utilizes pre-trained VLMs such as CLIP [40] for zero-shot 3D model generation [16, 19, 43], but with simple shape and texture styles. Subsequently, generative diffusion models equipped with advanced neural field representations [54] have demonstrated remarkable capabilities in high-fidelity 3D generation tasks via 2D-to-3D lifting, encompassing text-to-3D [15, 24, 27, 31, 37, 48, 52, 53, 57] and image-to-3D domains [26, 29, 38, 48, 57]. Notably, DreamFusion [37] and SJC [52] introduced the Score Distillation Sampling (SDS) optimization scheme by distilling 2D diffusion knowledge,
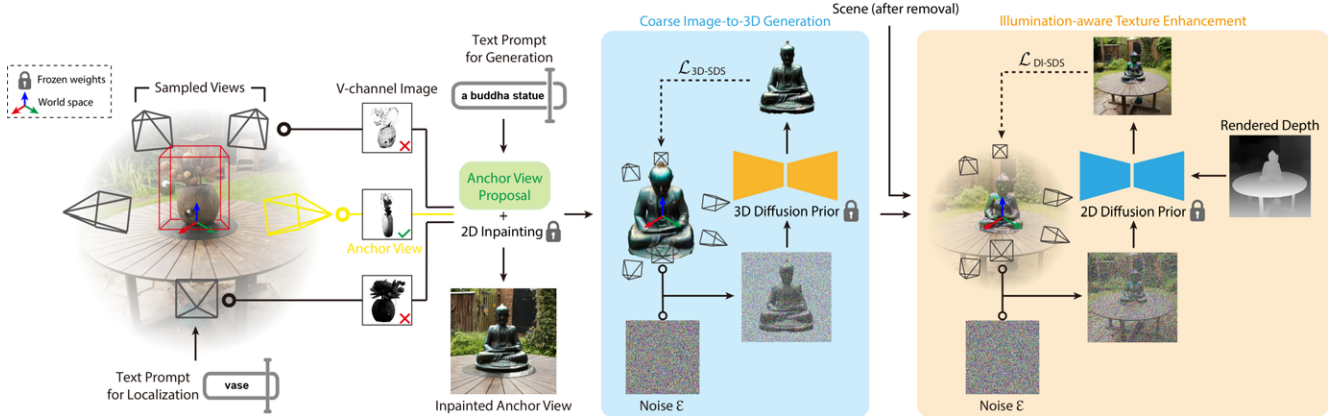
Figure 2. **Overview of Pipeline**. We propose a pipeline for localized scene editing, the new objects match the global illumination in the original scene. Our method uses text prompts to guide localization and generation, enabling us to synthesize and insert an object that is aware of the surrounding lighting. We introduce Anchor View Proposal to find a representative view of the illumination, which enables a two-step 3D lifting pipeline. Specifically, we propose illumination-aware Texture Enhancement with Depth-guided Inpainting Score Distillation Sampling (DI-SDS), which leverages contextual inpainting capability of large diffusion model. This enhancement step not only improves delicate texture details and motifs but also preserves global lighting conditions during optimization.

yielding encouraging text-to-3D results. The follow-up works further extended SDS to resolve issues such as low resolution and over-saturation. For example, Magic3D [24] employs a coarse-to-fine strategy using image and latent diffusion priors for high-quality generation, Prolific-Dreamer [53] addresses issues of over-saturation, over-smoothing, and low diversity through variational score distillation (VSD), and D-SDS [15] enhances SDS by incorporating score debiasing and prompt debiasing mechanisms. In parallel, with a single image input, Zero123 [27] excels in generating realistic novel perspectives via a view-dependent diffusion model, while SyncDreamer [29] synthesizes view-consistent novel perspectives without relying on neural fields through an intermediate synchronization statistics approach. With recent advanced 3D Gaussian Splatting (3DGS) [17] scene representation featuring swift training and rendering speed, [7, 48] increase generation speed and achieve text-to-3D generation with delicate texture details. These works primarily focus on single-object generation neglecting background environment interactions, such as ambient illumination and atmosphere. In contrast, our approach incorporates background effect into the 3D generation process with holistic contextual harmony between generation and background scene.

2.2. 3D Scene Editing

The rising popularity of implicit 3D representations, particularly radiance fields [17,33], has increased interest in 3D scene geometry and texture editing [1,10,25,28,35,49,58]. Advances in this field utilizes pre-trained vision, language and generative models [4,22,40,41,46], and achieves either direct modifications of entire scene [9,14,50] or intricate

manipulations and compositions of scene elements [1, 12, 20, 32, 51, 58]. Given a textual editing instruction, NeRF-Art [50] streamlines the modification of simple scenes by combining a foundational NeRF model with a style-oriented NeRF steered by CLIP priors. To progressively edit a NeRF scene, Instruct-NeRF2NeRF [14] alternatively updates training multi-views by Instruct-Pix2Pix [3] and then finetunes NeRF scene with updated training views. Later, to address the issue of multi-view inconsistent editing, ViCA-NeRF [9] enhances this process by integrating multi-view awareness with a blending mechanism. For more intricate scene alterations, SINE [1] advances image-guided 3D semantic editing by using an editing field informed by geometry and semantic vision priors [4, 8] to infuse detailed geometric and textural shifts. Meanwhile, some 3D editing works [32,44,45,47,58] focus on precise local changes with 3D segmentation and semantic editing. Within a target region in a NeRF scene, [47] achieves both geometry and texture alterations, and [44, 45] support object addition. DreamEditor [58] polishes local editing with a mesh-based field distilling from NeRF that can edit specific explicit segments within the 3D landscape. Concurrent works [6,11] employ 3DGS for swift text-guided scene editing. Compared to these editing works, we resolve the problem of object generation and replacement with dramatic geometry changes and background alignment.

3. Preliminaries

3.1. Score Distillation Sampling

Text-guided 3D generation has attracted increasing research interest and demonstrated significant progress by op-

timizing a 3D representation θ using a 2D pre-trained image diffusion prior ϵ_ϕ based on Score Distillation Sampling (SDS), as originally proposed in DreamFusion [37]. The diffusion model ϕ is pre-trained to predict sampled noise $\epsilon_\phi(x_t; t, y)$ at timestep t of the noisy image x_t, conditioned on text embeddings y. By rendering a random view by a differentiable renderer $g(\cdot)$, the aim is to minimize the discrepancy between the rendered image $\mathbf{x} = g(\theta)$ and the diffusion model distribution through the SDS loss. The scene parameterized by θ can be updated by computing the gradient derived from SDS loss:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, \mathbf{x}) = \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_\phi(x_t; t, y) - \epsilon)\frac{\partial \mathbf{x}}{\partial \theta}], \quad (1)$$

where $w(t)$ is a weighting function, ϵ is a Gaussian noise. Our method adapted SDS loss to incorporate the 3D-aware diffusion prior and a customized inpainting diffusion prior, compatible with 3D Gaussian Splatting representations.

3.2. 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [17] represents the scene with a set of anisotropic 3D Gaussians $\mathbf{G}(x)$ which can be efficiently projected into 2D Gaussians $\mathbf{G}'(x')$ for rendering. The differential rasterization process makes training and rendering faster than previous methods [2, 34]. By optimizing the multivariate Gaussians parameterized by position μ, anisotropic covariance $\Sigma = RSS^T R^T$, opacity α_i and color c_i information via posed multi-view images, 3DGS achieves an accurate representation of the scene with accelerated training and rendering times. The color $\mathbf{C}(x')$ and depth $\mathbf{D}(x')$ of each pixel x' in the image can be obtained via point-based volumetric rendering:

$$\begin{aligned} \mathbf{C}(x') &= \sum_{i \in \mathcal{N}} c_i \sigma_i \prod_{j=1}^{i-1}(1 - \sigma_j), \\ \mathbf{D}(x') &= \sum_{i \in \mathcal{N}} d_i \sigma_i \prod_{j=1}^{i-1}(1 - \sigma_j), \sigma_i = \alpha_i \mathbf{G}'(x') \end{aligned} \quad (2)$$

where $\mathcal{N}$ is ordered points overlapping the pixel, and d_i is the distance between 3D Gaussian center and camera center.

4. Method

Given an unbounded 3D Gaussian Splatting (3DGS) representation [17] scene, our objective is to perform text-guided localized 3D scene editing, specifically object insertion or replacement. We first leverage an off-the-shelf method [39] to localize a 3D bounding box as the target editing region. Then, we sample and render azimuth camera viewpoints around the bounding box to feed into our proposed Anchor View Proposal (AVP) module (Sec. 4.1), which votes for a view to be used in the 3D lifting step.

Next, we take a user-specified text prompt $\hat{y}$ as guidance to inpaint the anchor view, obtaining an image $\mathbf{x}_{inpaint}$. The object foreground of $\mathbf{x}_{inpaint}$, $\mathbf{x}_{inpaint}^{fore}$, is segmented out by [21] and fed into our coarse-to-fine 3D generation and texture enhancement pipeline (Sec. 4.2). In experiments, we assume the generation objects are opaque. The overview of our pipeline is illustrated in Fig. 2.

4.1. Anchor View Proposal

Existing multi-view diffusion models [27, 55] demonstrate the ability to generate consistent illumination across multi-views given the baked-in lighting in input single-view image. Inspired by this observation, we aim to identify an anchor viewpoint around the bounding box that includes the strongest illumination cues, such as shadow and highlight. To this end, we propose an Anchor View Proposal (AVP) algorithm to reliably select a well-conditioned view from rendering of N_{anc} surrounding viewpoints looking at the center of bounding box. Practically, we define a camera trajectory to render around the object and ensure capture of environmental illumination and no occlusion. The step may need certain amount of manual tuning such as distance to the object and elevation angle if the object in contact with a surface such as a table. The ideal proposed view should contain the most significant difference in illumination between the left-right, top-bottom, or diagonal halves of the rendered images (see evaluation in Sec. 5.2 and Sec. 5.4). This is achieved by rotating each rendered image by a predefined set of Azimuth angles and selecting the rotation that places majority of brighter pixels to the left. For robust lighting estimation from images, we convert rendered RGB images to HSV color space. This design choice is enlightened by the fact that V is the distance along the axis that turns black into non-black color, also known as "brightness". Therefore, computation of V-channel image estimates amount of light that lit the surface. In scenarios where the majority of illumination is predominantly cast in the top-bottom direction, our algorithm tends to output an arbitrary view. This behavior aligns with our requirement, as we sample viewpoints in azimuth and any view is considered equally suitable. Pseudocode of AVP algorithm is provided in the supplementary material.

4.2. Context-aware Coarse-to-Fine 3D Generation

After acquiring the anchor view, we use an off-the-shelf depth-conditioned diffusion model [56] to inpaint the mask of bounding box projection, according to the generation text prompt $\hat{y}$. The foreground served as input to the next 3D lifting pipeline can be extracted by the segmentation model [21]. For efficient and robust 3D generation with contextual illumination awareness, we propose to conduct a coarse but swift image-to-3D generation via 3D-aware diffusion prior, and an illumination-aware texture enhance-

ment step to further realize high-quality 3D lifting result. Note that after removal (if any), we leave the scene points untouched and only optimize new generated points.

4.2.1 Coarse Image-to-3D Generation

Recent efficient multi-view diffusion models [27] were pre-trained on large-scale 3D objects with backed lighting. Given the lighting-heuristic inpainted anchor view, in the coarse step, the multi-view diffusion model can faithfully lift the inpainted image to 3D with multi-view consistency.

To achieve reliable 3D lifting in 3DGS representation, we adopt compact-based densification and pruning strategy proposed in [7]. Instead of initializing 3D Gaussians from Point-E [36], we initialize them in a solid sphere, with opacity in a normal distribution proportional to the center distances. This strategy often performs better in our coarse step, since the inpainted conditioning image may not align with the generation modality in Point-E, leading to inconsistent points with the input image and longer training time to converge. Still, some initial points may remain as floaters and are hard to be pruned later. Therefore, we significantly lower initial opacity and colors of 3D Gaussians.

During object optimization by randomly sampled views, we aim to minimize the objective function:

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{3D-SDS}\mathcal{L}_{3D-SDS}, \quad (3)$$

where $\mathcal{L}_{rgb}$ is the mean squared error (MSE) between the foreground anchor view image $\mathbf{x}_{inpaint}^{fore}$ and rendered image at the same viewpoint, $\mathcal{L}_{mask}$ is MSE loss between ground-truth mask extracted from $\mathbf{x}_{inpaint}^{fore}$ and predicted opacity image $\mathbf{m}_{anchor}$, $\mathcal{L}_{3D-SDS}$ is score distillation loss (Eqn. 1) from 3D-aware diffusion prior [27] for the sample view. The weight of reference view RGB (λ_{rgb}) increases as training proceeds. We compute $\mathbf{m}_{anchor}$ by volumetric rendering equation:

$$\mathbf{m}_{anchor} = \sum_{i \in \mathcal{N}} \mathbf{1} \cdot \sigma_i \prod_{j=1}^{i-1}(1 - \sigma_j). \quad (4)$$

With $\mathcal{L}_{mask}$, we encourage foreground points to be fully opaque and background points to be transparent.

4.2.2 Illumination-aware Texture Enhancement

With the first coarse generation step, the synthesized object has a multi-view consistent shape and appearance, but lacks subtle texture details and high resolution required by photo-realism. Therefore, we propose a contextually illumination-aware Texture Enhancement step, which not only enriches geometry and texture details, but also preserves multi-view lighting conditions.

While we observe pre-trained diffusion model with SDS (Eqn. 1) can greatly enhance texture and geometry details,

it does not receive background information from the scene. Inspired by 2D inpainting, we propose Depth-guided Inpainting Score Distillation Sampling (DI-SDS), where intermediate output features of depth-guided ControlNet copy network $\mathcal{C}_{depth}$, projected bounding box mask and masked image are fed into the diffusion model for conditioned generation. Specifically, we first feed colored image latents x_t, timestep t, text prompt embeddings y and depth image $\mathbf{d}$ into ControlNet layers $\mathcal{C}_{depth}$ to obtain high-level residual block output samples D and middle-level output samples M:

$$(D, M) = \mathcal{C}_{depth}(x_t; t, y, \mathbf{d}), \quad (5)$$

where depth image $\mathbf{d} = 1 - d$ (i.e., inversed depth), and d is derived from volumetric rendering process (Eqn. 2) and scaled into range $[0, 1]$.

Then, we concatenate bounding box mask m and masked image latents m_l with colored image latents, and feed them additionally into 2D diffusion model ϕ with D, M:

$$\nabla_\theta \mathcal{L}_{DI-SDS}(\phi, \mathbf{x}) =$$
$$\mathbb{E}_{t,\epsilon}[w(t)(\epsilon_\phi(x_t, m, m_l; t, y, D, M) - \epsilon)\frac{\partial \mathbf{x}}{\partial \theta}], \quad (6)$$

where θ is 3DGS scene parameters, $\epsilon_\phi(\cdot)$ is predicted sampled noise, ϵ is a Gaussian noise, $\mathbf{x}$ is rendered RGB image.

We accommodate classifier-free guidance by computing predicted noise with the following equation:

$$\hat{\epsilon_\phi}(x_t, m, m_l; t, y, \mathbf{d}) = \epsilon_\phi(x_t, m, m_l; t, y, \mathbf{d})$$
$$+ s \cdot (\epsilon_\phi(x_t, m, m_l; t, y, \mathbf{d}) - \epsilon_\phi(x_t; t)), \quad (7)$$

where s is guidance scale specified in Sec. 5.1.

The text prompt guiding generation $\hat{y}$ is embedded into view-conditioned text embeddings y, for example, "a standing pineapple, photorealistic, front view", to prevent Janus problem and provide multi-view information. Since we optimize both object geometry and texture through finetuning, we lower the learning rate in geometry with a focus on smooth texture refinement. During optimization, we merge Gaussian splats of background color field and generated object splats, and render randomly sampled views following strategy in [13]. In Texture Enhancement step, we disable $\mathcal{L}_{3D-SDS}$ because it prevents details generated by 2D inpainting diffusion model that does not share the same prior distribution. Therefore, the loss function of this step becomes

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{DI-SDS}\mathcal{L}_{DI-SDS}. \quad (8)$$

Providing flexible controllability for users, this Texture Enhancement step can be balanced by fewer optimization iterations in some cases when simple geometry and texture patterns are sufficiently satisfactory. For example, the monkey statue is expected to have smooth skin in Fig. 6.
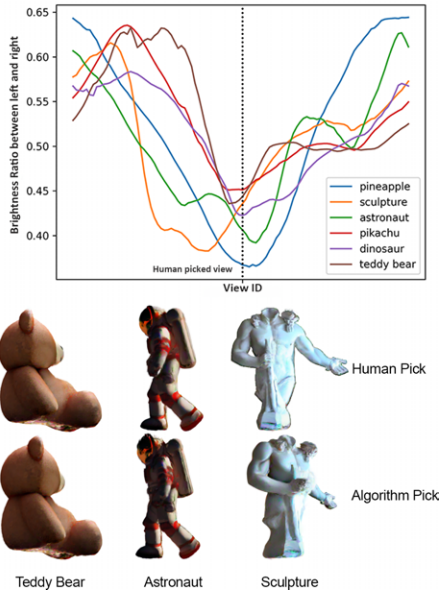
Figure 3. Evaluation on Anchor View Proposal algorithm. We evaluate the robustness of our Anchor View Proposal (AVP) algorithm by comparing view angle differences between human and algorithm proposals on six objects. The top graph shows the Value-channel (in HSV space) brightness ratio (range: $[0, 1]$) for 100 continuous azimuth views, where a lower value indicates the left side is darker. In this context, AVP selects views with the lowest difference ratio, closely matching human selections. The bottom examples further demonstrate AVP's effectiveness in selecting high contrast views.

5. Experiment

5.1. Implementation Details

Datasets. To evaluate performance in diverse scenes, we use LERF [18], MipNeRF360 [2] and two self-captured datasets where we set up directional light sources. We focus on scenes with challenging illumination that may include multiple objects to test robustness of our pipeline.

Pre-trained Models. We use released Stable Zero123 model in coarse step, ControlNet Depth version [56] in 2D inpainting and Texture Enhancement step, and Stable Diffusion [42] as 2D diffusion model. Specifically, we leveraged Stable Diffusion Inpainting model by Runway as 2D diffusion model in Texture Enhancement. For view conditioned text embedding, we used Stable Diffusion v1.5 by Runway. Forwarding through ControlNet Depth and Stable Diffusion together provides predicted noise residual. Parameters of all diffusion models are frozen in training.

5.2. Qualitative Results

We mix and match text prompt with scenes for generation and render multi-view images for qualitative evaluation in Fig. 6. In the figure, we show the rendering of input scene
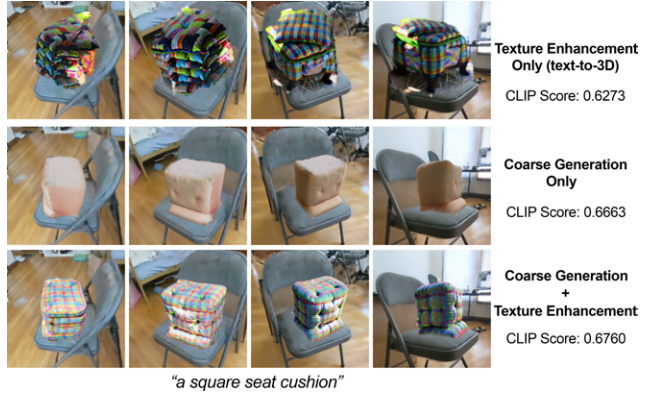

Figure 4. Ablation study on Contextual 3D Lifting. We show novel views of localized object editing when we skip coarse step, skip Texture Enhancement step, and apply full pipeline. Coarse step aims to provide initial prior and scene illumination in texture. Texture Enhancement aims to generate more details in geometry and texture.
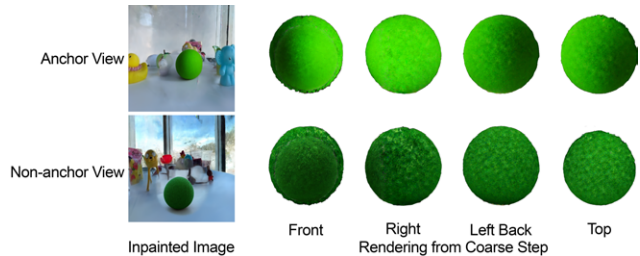

Figure 5. Ablation study on Anchor View Proposal in terms of influence to image-to-3D result. Top row: proposed anchor view. Bottom row: non-anchor view. We show inpainted images with the same generator seed in the left column and multi-view rendering of 3D lifted object. Conditioning view that does not reflect scene illumination fails to generate texture that fits into the scene from other viewpoints.

under anchor view with 3D bounding box visualized, and 2D inpainted object under proposed anchor view. Then, we show four multi-view images rendered from output scene. Our result demonstrates detailed texture and faithful respect to scene illumination.

Evaluation on Anchor View Proposal algorithm. We selected six bounding boxes and their associated objects with strong illumination, then generated multi-view images as input into AVP algorithm. Specifically, we sample 100 object-centered viewpoints along azimuth with every two nearby views are 3.6 degrees apart. Then, we had human manually pick the "best" view with the strongest highlight or shadow on left or right half depending on requirements. We graph the offset in viewpoint indices as selected by both the human and our algorithm, and present the three pairs of views with the highest offsets in Fig. 3. In the plot, we observe that the difference in brightness between left and right half strongly implies the optimal viewpoint to pick. In
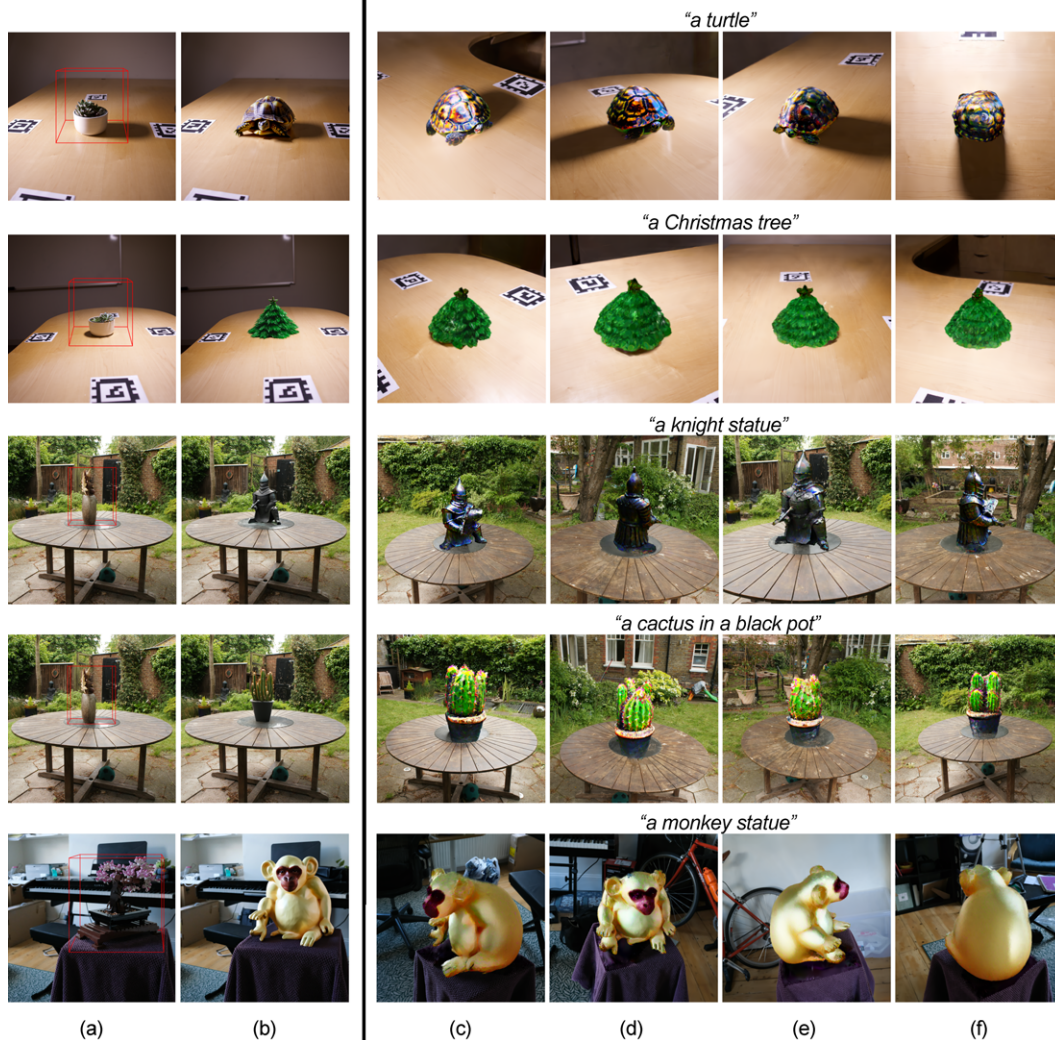
Figure 6. Qualitative Results. (a) is input pre-trained 3D Gaussian Splatting scene with bounding box marked in red. (b) is anchor view image with inpainted object according to text prompt on top of each row. (c-f) are multi-view rendering of output scenes. More results are presented in supplementary material.

the case with the highest offset (i.e., the white sculpture), both the human-picked and the algorithm-proposed views are good enough as conditioning viewpoint.

5.3. Comparison

Qualitative Comparison. We compare our method against Vica-NeRF [9] and GaussianEditor [6] by presenting qualitative result in Fig. 7. While we additionally provide localization information in text prompt to Vica-NeRF, it is unable to edit the region of interest or preserve background color. Same as ours, GaussianEditor performs segmentation via SAM and 3D lifting from 2D inpainted image. However, GaussianEditor neither includes a mechanism to pick views containing illumination information, nor uses background in multi-view optimization [30]. There-

fore, it suffers from negligence to the scene illumination such as Pikachu and astronaut cases in Fig. 7.

Quantitative Comparison. We conducted a user study on the results and collected statistics of user's preference. In the questionnaire, we put outputs of different methods in random order and ask participants which one is the most consistent with the text prompt and looks the most realistic within the scene. Statistics show that 0% users prefer Vica-NeRF, 29.49% prefer GaussianEditor and 70.51% prefer ours.

5.4. Ablation Study

Effect on Coarse-to-Fine 3D Generation. In Fig. 4, we show multi-view rendering of a generated square cushion on a chair when we use texture enhancement only, coarse

Figure 7. Comparison. We show visual results from ours, GaussianEditor [6] and Vica-NeRF [9], along with a reference image of original scene. Specifically, we provide 1) longer text prompt (with words in square bracket) on top of each row containing localization information to Vica-NeRF; 2) shorter text prompt (without words in square brackets) and pre-determined bounding box to ours and GaussianEditor.

step only, and full pipeline. Without coarse step, the generated object suffers from inconsistency in multi-view and weak fitting to scene illumination. Without Texture Enhancement, we observe consistent geometry, smooth texture, and faithful lifting from anchor view image; however, more details in texture can be added while generation is still aligned with text prompt. With both steps, we observe a significant amount of detail added to both geometry and texture while highlight and shadow cast on the surface still respect the scene. We provide average CLIP score of multi-view rendering following the same way in quantitative comparison (Sec. 5.3).

Effect on Anchor View Proposal. In Fig. 5, we show multi-view rendering of coarse step output when we inpaint the proposed anchor view and another non-anchor view with same inpainting seed and input text prompt "a green ball". With the proposed inpainted anchor view, our coarse step and 3D-aware diffusion prior are robust enough to generate multi-views in contextually consistent illumination. The anchor view is more suitable for 3D photorealism, as it implies heuristic illumination information.

6. Conclusion

In this paper, we proposed an end-to-end pipeline to address text-guided localized editing of 3D scenes, focusing on generation of objects that naturally fit into the scene with illumination awareness. We observed that pre-trained 2D inpainting diffusion model is able to generate objects with texture consistent with scene illumination. Inspired by this, we proposed an Anchor View Proposal algorithm to find a viewpoint, under which we can obtain a conditioning image that contains the most contextual lighting information. We then proposed context-aware coarse-to-fine 3D generation pipeline that faithfully lifts conditioning image into 3D via a coarse step, and then adds details into geometry and texture via a Texture Enhancement step. To resolve the bottleneck that previous SDS does not support optimization with rendering background, we derived DI-SDS from depth inpainting ControlNet to supervise the Texture Enhancement. Experiments demonstrate that our pipeline successfully distills directional illumination from the input scene. Especially, we want to inspire the field to explore derivation and extension to SDS in conditioned generation and editing tasks.

Acknowledgement

References

[1] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20919–20929, 2023. 2, 3

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 4, 6

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, June 2023. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[5] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 2

[6] 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*, 2024. 3, 7, 8

[7] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3, 5

[8] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *IEEE Computer Vision and Pattern Recognition*, 2021. 3

[9] Jiahua Dong and Yu-Xiong Wang. ViCA-neRF: View-consistency-aware 3d editing of neural radiance fields. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 7, 8

[10] Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *European Conference on Computer Vision*, 2022. 3

[11] Jiemin Fang, Junjie Wang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *CVPR*, 2024. 2, 3

[12] Ori Gordon, Omri Avrahami, and Dani Lischinski. Blended-nerf: Zero-shot object generation and blending in existing neural radiance fields. In *International Conference on Computer Vision AI3DCC Workshop*, 2023. 3

[13] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 5

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

[15] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for view-consistent text-to-3d generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[16] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 2

[17] Kerbl, Bernhard and Kopanas, Georgios, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 2, 3, 4

[18] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 6

[19] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Popa Tiberiu. Clip-mesh: Generating textured meshes from text using pretrained image-text models. *SIGGRAPH Asia 2022 Conference Papers*, December 2022. 2

[20] Hyunsu Kim, Gayoung Lee, Yunjey Choi, Jin-Hwa Kim, and Jun-Yan Zhu. 3d-aware blending with generative nerfs. In *ICCV*, 2023. 3

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 4

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 3

[23] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. *arXiv preprint arXiv:2311.16473*, 2023. 2

[24] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[25] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8338–8348, 2023. 3

[26] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[27] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2, 3, 4, 5

[28] 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. 3

[29] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[30] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion, 2023. 7

[31] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. Att3d: Amortized text-to-3d object synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[32] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. OBJECT 3DIT: Language-guided 3d-aware image editing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[33] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[34] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 4

[35] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2022*, 2022. 3

[36] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts, 2022. 5

[37] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4

[38] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors, 2023. 2

[39] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *CVPR*, 2024. 4

[40] 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. 2, 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 6

[43] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 2

[44] Mohamad Shahbazi, Liesbeth Claessens, Michael Niemeyer, Edo Collins, Alessio Tonioni, Luc Van Gool, and Federico Tombari. Inserf: Text-driven generative object insertion in neural 3d scenes. *Arxiv*, 2024. 3

[45] Ka Chun Shum, Jaeyeon Kim, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Language-driven object fusion into neural radiance fields with pose-conditioned dataset updates. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[47] Hyeonseop Song, Seokhun Choi, Hoseok Do, Chul Lee, and Taehyeong Kim. Blending-nerf: Text-driven localized editing in neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14383–14393, 2023. 2, 3

[48] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[49] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3825–3834, 2021. 3

[50] 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. 3

[51] Dongqing Wang, Tong Zhang, Alaa Abboud, and Sabine Süsstrunk. Inpaintnerf360: Text-guided 3d inpainting on unbounded neural radiance fields, 2023. 2, 3

[52] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting

pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[53] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3

[54] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022. 2

[55] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. In *3DV*, 2024. 4

[56] 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*, pages 3836–3847, 2023. 1, 2, 4, 6

[57] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang YU, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[58] Jingyu Zhuang, Chen Wang, Liang Lin, Lingjie Liu, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 2, 3