Personalize Your Gaussian: Consistent 3D Scene Personalization from a Single Image

Yuxuan Wang[1], **Xuanyu Yi**[1], **Qingshan Xu**[1], **Yuan Zhou**[1], **Long Chen**[2], **Hanwang Zhang**[1]
[1]Nanyang Technological University
[2]The Hong Kong University of Science and Technology

Abstract

Personalizing 3D scenes from a single reference image enables intuitive user-guided editing, which requires achieving both multi-view consistency across viewpoints and referential consistency with the input image. However, these goals are particularly challenging due to the viewpoint bias caused by the limited perspective provided in a single image. Lacking the mechanisms to effectively expand reference information beyond the original view, existing methods of image-conditioned 3DGS personalization often suffer from this viewpoint bias and struggle to produce consistent results. Therefore, in this paper, we present **Consistent Personalization for 3D Gaussian Splatting (CP-GS)**, a framework that progressively propagates the single-view reference appearance to novel perspectives. In particular, CP-GS integrates pre-trained image-to-3D generation and iterative LoRA fine-tuning to extract and extend the reference appearance, and finally produces faithful multi-view guidance images and the personalized 3DGS outputs through a view-consistent generation process guided by geometric cues. Extensive experiments on real-world scenes show that our CP-GS effectively mitigates the viewpoint bias, achieving high-quality image-conditioned 3DGS personalization that significantly outperforms existing methods.

1 Introduction

In the evolving field of 3D computer vision, user-friendly 3D editing has attracted growing attention as a key research focus (Chen et al. 2023b; Haque et al. 2023; Karim et al. 2023; Dong and Wang 2024; Song et al. 2023; Chen, Laina, and Vedaldi 2024; Wang et al. 2024; Lee et al. 2024; Dihlmann, Engelhardt, and Lensch 2024; Fang et al. 2023). Among recent advances, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) has emerged as a groundbreaking 3D representation, offering an explicit and efficient structure that supports local manipulation and rendering in real time. Building up the 3DGS representation, we focus on a practical and intuitive form of user interaction—personalizing a 3DGS scene using only a single-view reference image—by editing a user-specified region to match the reference appearance. To make it clear, as illustrated in Figure 1, given a reference image depicting a unique brown ***panda***, our goal is to modify a user-specific ***bear*** region in the scene to a ***panda*** that aligns with the appearance in the image. This task enables intuitive 3D customization from a single reference image, supporting applications such as personalized avatars in virtual reality and assets stylization in interactive environments.

With the advent of large-scale pre-trained 2D diffusion models (Rombach et al. 2022; Podell et al. 2023; Peebles and Xie 2023), recent 3DGS editing methods (Chen, Laina, and Vedaldi 2024; Wang et al. 2024; Fang et al. 2023; Wu et al. 2024) have predominately leveraged image generation models to produce pseudo-images as editing guidance that supervise the fine-tuning of 3DGS scenes. In this paradigm, the task of image-conditioned personalization requires two key consistencies in the guidance images: (1) **referential consistency** with the visual appearance of the reference image and (2) **multi-view consistency** across different perspectives to prevent conflicting guidance. However, achieving these consistencies remains a significant challenge for existing approaches (Zhuang et al. 2024) conditioned on a single reference image. As illustrated in Figure 2, prior methods typically adapt their image generation models directly to the single reference view, often overfitting and misprojecting appearance features in the reference image onto unrelated viewpoints. This leads to distorted appearances and severe multi-view inconsistencies in the editing guidance, ultimately resulting in noticeable artifacts in the final 3D output.

We argue that the core challenge lies in the viewpoint bias introduced by the limited perspective of a single reference image, where the image model lacks sufficient information to infer appearances under novel viewpoints that are far from the reference. As a result, the model is often biased towards the reference view, making existing methods struggle to produce consistent multi-view editing guidance. Therefore, in this paper, we propose **Consistent Personalization for 3DGS (CP-GS)**, a high-quality personalization framework that addresses the viewpoint bias by progressively propagating the reference appearance to novel perspectives. As illustrated in Figure 2, using a coarse guidance as structural priors to establish viewpoint cues for consistent appearance generation, CP-GS operates in a coarse-to-fine manner with three stages: (1) Coarse guidance generation to initialize geometry and propagate rough appearance. (2) Iterative LoRA fine-tuning to extract and extend fine-grained reference details. (3) View-consistent generation that leverages the coarse guidance and trained LoRA to produce the final guidance images, which

Figure 1: Given a source 3DGS scene and a single reference image, CP-GS enables high-quality personalization by editing a user-specified region (e.g., the *bear*, *man's eye*, *man's face*, *bench-top*, *entire scene*) to match the reference appearance, supporting object replacement, adding new objects, and stylization of existing objects.

are used to fine-tune and produce the 3DGS output.

In the first stage, we establish a coarse guidance that serves as a structural prior, enabling the initial propagation of reference appearance into a coarse, view-consistent 3D representation. Specifically, we employ a pre-trained image-to-3D generation model (Xiang et al. 2024) to produce a geometry-consistent contour with a rough texture estimate, which is integrated into the target location in the scene. As shown in Figure 2, although the resulting textures are often unrealistic due to the domain gap between the real-world reference image and the CGI-style pre-training data (Deitke et al. 2023b,a) of image-to-3D model, it captures structural geometry and a rough yet view-consistent appearance.

To recover fine-grained reference appearance while avoiding the viewpoint bias, the second stage draws inspiration from (Raj et al. 2023), which proposed that diffusion models adapted to a single image hold the potential to generate novel neighboring views around the reference. Therefore, in this stage, we propose an iterative LoRA fine-tuning strategy that gradually extracts and propagates reference appearance to novel viewpoints. In each iteration, we translate novel-view renderings of the coarse guidance using the current trained LoRA model and select one well-aligned result—identified via our designed scoring mechanism based on dense feature matching (Edstedt et al. 2024)—to augment the training set for the next round fine-tuning.

Leveraging the coarse guidance and the trained LoRA, we employ a pre-trained flow-based model (Labs 2024) in the last stage to generate the final guidance images. We begin by applying rapid rectified-flow inversion (Rout et al. 2024) to convert renderings of the coarse guidance into noisy latents, which are passed to the Flow Transformer and serve as the starting point for generation, conditioned on the depth maps of the coarse guidance. To further reduce the multi-view inconsistency arising from viewpoint variance, we introduce an epipolar-constrained token replacement strategy that aggregates visual features across views using geometric correspondences, improving overall multi-view coherence.

As illustrated in Figure 1, by progressively propagating the single-view reference appearance in a coarse-to-fine manner, CP-GS effectively addresses the challenges posed by the viewpoint bias, resulting in superior visual quality in personalized 3DGS results. Comprehensive evaluations across diverse real-world scenes demonstrate that our CP-GS successfully address the artifacts caused by limited reference perspectives and outperforms state-of-the-art methods in both qualitative and quantitative comparisons. Based on the above, our contributions can be summarized in three aspects:

- We identify the viewpoint bias caused by limited reference perspective information as the crux of referential and multi-view inconsistencies in previous single-view 3D personalization methods.
- To mitigate the viewpoint bias, we propose a coarse-to-fine appearance propagation framework that progressively expands the single-view reference appearance to novel perspectives, generating guidance images with faithful referential consistency and strong multi-view consistency.
- We validate CP-GS through extensive experiments on various real-world scenes, demonstrating its superior performance over previous 3DGS personalization and editing methods in both qualitative and quantitative evaluations.

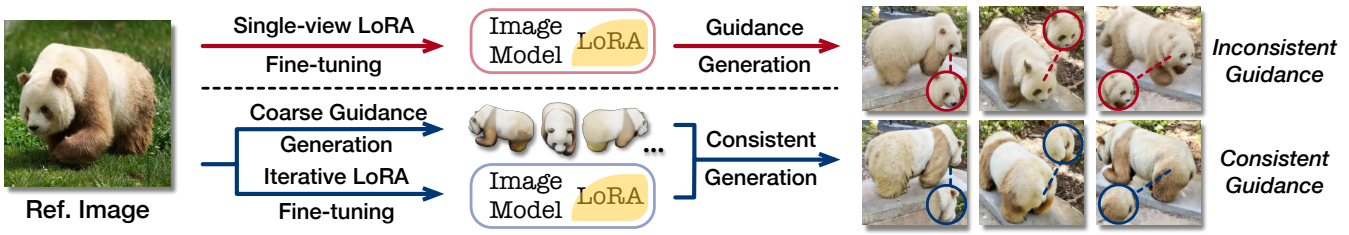

Figure 2: ***red***: Previous methods suffer from viewpoint bias and produce distorted editing guidance, where the resulting panda shows significant viewpoint overfitting, leading to both the referential and multi-view inconsistencies. ***blue***: By progressively propagating reference to novel views, CP-GS mitigates the bias and achieves both consistencies in the guidance images.

2 Related Works

Image-guided 2D Customization. Given a set of reference images, the task of 2D customization aims to edit a source image or generate a new image under the guidance of the reference, where customization methods (Gal et al. 2022a; Ruiz et al. 2023; Liu et al. 2023) typically optimize a special token or use LoRA-based adaptation to capture the appearance of the reference images. Built on this strategy, early methods (Gal et al. 2022a; Ruiz et al. 2023; Liu et al. 2023; Choi et al. 2023; Yang et al. 2023) rely on multiple reference images to construct the novel content through test-time fine-tuning (TTF). Subsequent works (Chen et al. 2023a; Li et al. 2023; Sohn et al. 2023; Avrahami et al. 2023) further improve the flexibility of this paradigm by training with a single reference image. Recently, leveraging large-scale image datasets (Schuhmann et al. 2022; Pexels 2024; Byeon et al. 2022; Changpinyo et al. 2021; Sharma et al. 2018), a line of work (Li, Li, and Hoi 2023; Chen et al. 2024a; Yuan et al. 2023; Chen et al. 2024b,c; Peng et al. 2024) has adopted pre-trained adaptation (PTA), which trains on large-scale paired data and bypasses fine-tuning during inference. While our iterative LoRA fine-tuning strategy builds on the test-time fine-tuning paradigm, the task of 3D personalization presents additional challenges beyond those in 2D customization, notably the need for multi-view consistency and mitigating the viewpoint bias to avoid conflict guidance.

Consistent 3D Field Editing. Early approaches for consistent 3D editing (Kamata et al. 2023; Park, Kwon, and Ye 2023; Yi et al. 2024; Zhuang et al. 2023; Koo, Park, and Sung 2023) predominantly rely on NeRF (Mildenhall et al. 2021) representations optimized via Score Distillation Sampling (SDS)-based techniques (Poole et al. 2022). Subsequent works (Chen et al. 2023b; Haque et al. 2023; Dong and Wang 2024) employ image-guided 3D editing by leveraging pre-trained 2D diffusion models to generate multi-view guidance images. Pioneered by (Chen et al. 2023b; Fang et al. 2023), recent methods integrate Gaussian Splatting (Kerbl et al. 2023) into 3D field editing due to its superior efficiency and controllability. More recently, a line of research (Chen, Laina, and Vedaldi 2024; Wang et al. 2024; Lee et al. 2024; Wu et al. 2024) has aimed to explicitly ensure multi-view consistency in the guidance images. VcEdit (Wang et al. 2024) introduces latent and attention map aggregation, while GaussCtrl (Wu et al. 2024) and DGE (Chen, Laina, and Vedaldi 2024) utilize cross-view extensive attention to harmonize the variations across views. However, all these methods are limited to simple text prompts condition and lack the ability of customized editing. The most relevant work with ours is TIP-Editor (Zhuang et al. 2024), which combines the LoRA fine-tuning and the SDS optimization to distill the reference content into 3D scene. However, it fails to consistently expand the reference appearance across views, often exhibits visual artifacts in the 3D outputs due to viewpoint bias.

3 Methodology

In this section, we present the **CP-GS** framework that personalize a 3DGS scene from a single-view reference image (Sec. 3.1), with the overall pipeline illustrated in Figure 3. We first employ a pre-trained image-to-3D model to construct a coarse guidance with a rough yet view-consistent reference appearance, serving as the initial step of our propagation (Sec. 3.2). To further extract and propagate fine-grained appearance details, we then introduce an iterative LoRA fine-tuning strategy that progressively expands the training views through image translation and selective augmentation (Sec. 3.3). Finally, we combine the coarse guidance and the trained LoRA within a pre-trained Flow model to generate multi-view consistent, reference-aligned guidance images, resulting in the final 3DGS output (Sec. 3.4).

3.1 Problem Definition

Given a source 3DGS scene $\mathcal{G}^{\text{src}}$ and a reference image $\mathcal{I}^{\text{ref}}$, the goal is to edit a user-specific region in $\mathcal{G}^{\text{src}}$ to the personalized $\mathcal{G}^{\text{edit}}$ that align with $\mathcal{I}^{\text{ref}}$. To achieve this, we adopt an image-guided paradigm that generates a set of multi-view personalized guidance images $\mathcal{I}^{\text{gde}}$ to supervise the transformation of $\mathcal{G}^{\text{src}}$ into $\mathcal{G}^{\text{edit}}$. For each view, we define an editing loss including a mean absolute error $\mathcal{L}_{\text{MAE}}$ and a perceptual loss $\mathcal{L}_{\text{LPIPS}}$ between the real-time rendering and corresponding guidance image. The final 3DGS model $\mathcal{G}^{\text{edit}}$ is then optimized by minimizing the editing loss across all views $\mathcal{V}$:

$$\mathcal{G}^{\text{edit}} = \underset{\mathcal{G}}{\text{argmin}} \sum_{v \in \mathcal{V}} (\lambda_1 \mathcal{L}_{\text{MAE}}(\mathcal{R}(\mathcal{G}, v),\, \mathcal{I}^{\text{gde}}) + \lambda_2 \mathcal{L}_{\text{LPIPS}}(\mathcal{R}(\mathcal{G}, v),\, \mathcal{I}^{\text{gde}})) \quad (1)$$

where $\mathcal{R}$ denotes the rendering function (Kerbl et al. 2023). This paradigm requires two key properties in the multi-view guidance images $\mathcal{I}^{\text{gde}}$: multi-view consistency across $\mathcal{V}$ to prevent optimization conflicts, and referential appearance

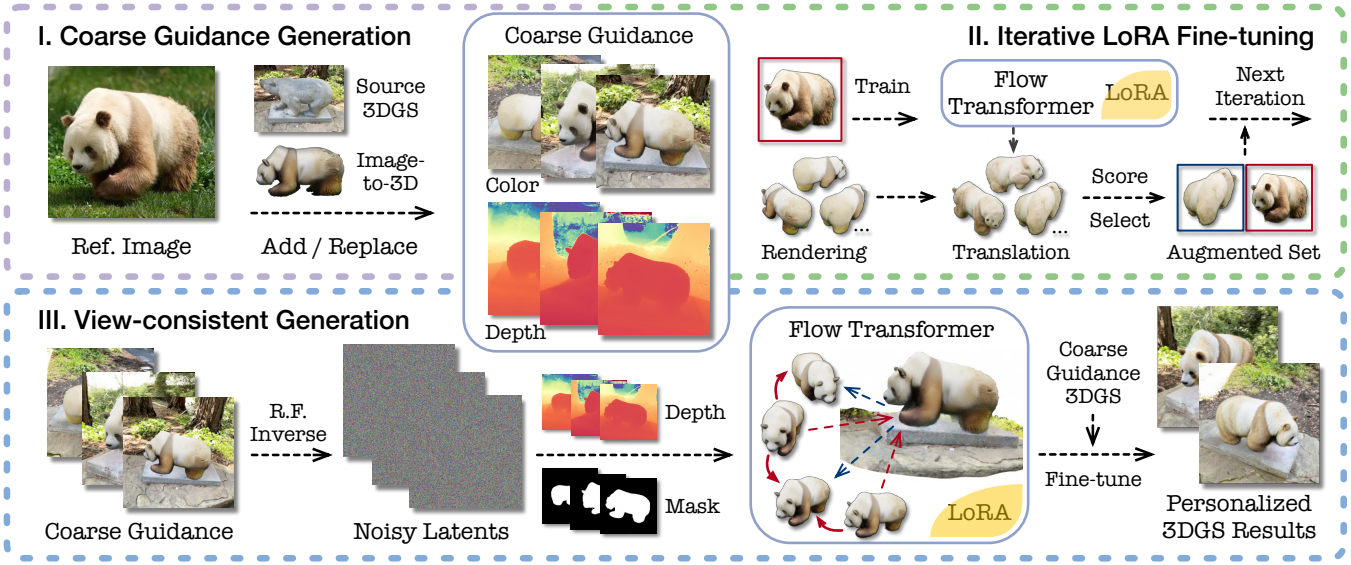

Figure 3: The pipeline of our **CP-GS** includes three stages: coarse guidance generation via a pre-trained image-to-3D model; iterative LoRA fine-tuning to extract and propagate detailed reference appearance; and view-consistent generation of guidance images to produce final 3DGS outputs, progressively propagating the single reference image to a multi-view representation.

consistency to the $\mathcal{I}^{\text{ref}}$ to fulfill the personalization objective. Our CP-GS is designed to explicitly ensure both consistency to achieve high-quality 3D personalization.

3.2 Coarse Guidance Generation

As noted in Sec. 1, the limited perspective of a single reference image fails to provide sufficient geometric and coherent appearance information for constructing a consistent multi-view representation. Therefore, in the first stage, we leverage an off-the-shelf image-to-3D generation model TRELLIS (Xiang et al. 2024) pre-trained on large-scale CGI-style 3D datasets (Deitke et al. 2023b,a) to produce a coarse guidance, expanding the reference into a rough yet multi-view consistent representation. As illustrated in the *top-left* of Figure 3, the reference image is fed into the pre-trained TRELLIS to generate the corresponding 3D rough asset, which is then integrated into the source scene to replace or augment the user-specific target region. Here we provide two integration modes: (1) Adding new content – the user provides a 3D bounding box specifying the object's position and scale; (2) Replacing existing content – the target bounding box is extracted from the existing content via PCA (Abdi and Williams 2010), and the generated asset is fitted accordingly. This coarse guidance provides a plausible 3D geometry and establishing a rough yet view-consistent appearance that serves as a structural prior for subsequent stages.

3.3 Iterative LoRA Fine-tuning

Due to the inevitable domain gap between the real-world reference image and the CGI-style datasets (Deitke et al. 2023b,a) used to train the image-to-3D model (Xiang et al. 2024), the generated assets of coarse guidance often exhibit unrealistic and rough appearance that lacks referential consistency. On the other hand, we observe that image generation model tend to overfit to the reference view under the single image setting, resulting in a strong viewpoint bias. To address these issues, we adopt an iterative LoRA fine-tuning strategy that retrieves a fine-grained appearance from the reference and progressively propagates it to novel views.

Specifically, we initialize the training set with the given single-view image, conducting the first iteration of LoRA fine-tuning using a prompt containing a learnable special token to encode the reference characteristics. Inspired by DreamBooth3D (Raj et al. 2023), which demonstrates that image generation models (Rombach et al. 2022) adapted to a single-view image can still synthesize neighboring views of the reference subject within a limited perspective range, we render the coarse guidance from multiple viewpoints and apply the fine-tuned model to translate the renderings toward the photorealistic reference target using the same prompt. Subsequently, we select one well-aligned translated image using a task-specific scoring mechanism and append it to the training set to augment the next round of fine-tuning.

We notice that designing such scoring mechanism is non-trivial, as it must avoid viewpoint-biased translations and redundant views that are already well covered by the training set, ensuring that each selection contributes effectively to our appearance propagation. Notably, as shown in Figure 4(a), both types of undesirable cases exhibit relatively high similarity to the training images: (1) redundant views, which are close to the training perspective, naturally share similar appearance; and (2) biased translations, which often inherit excessive training-view features due to overfitting, also tend to exhibit higher similarity to training images than the well-aligned novel-view results. Therefore, we identify the well-aligned result as the one with minimal overall similarity to the training set, measured via dense feature matching using the pre-trained RoMa model (Edstedt et al. 2024). De-

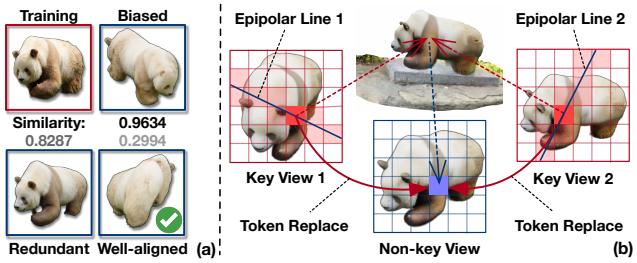

Figure 4: **(a)** Visualization of the translated results and the similarities in our scoring mechanism. **(b)** Illustration of the proposed epipolar-constrained token replacement strategy.

noting $\mathcal{I}_t^{\text{train}}$ the training image set and $\mathcal{I}_t^{\text{trans}}$ the translations at iteration t, our scoring and selection are formulated as:

$$\mathcal{I}_{t+1}^{\text{train}} = \mathcal{I}_t^{\text{train}} \cup \{\arg\min_i \sum_j \mathbf{S}_{\text{RoMa}}(I_i^{\text{trans}}, I_j^{\text{train}})\},$$
$$\text{where } I_i^{\text{trans}} \in \mathcal{I}_t^{\text{trans}}, \; I_j^{\text{train}} \in \mathcal{I}_t^{\text{train}} \quad (2)$$

where $\mathbf{S}_{\text{RoMa}}(\cdot)$ denotes the similarity computed by the RoMa model. Leveraging the neighboring-view expansion capability of the trained LoRA module, our iterative fine-tuning strategy effectively propagates single-view reference details to novel perspectives and alleviates the viewpoint bias, enabling the model to learn fine-grained appearance with both multi-view and reference consistency.

3.4 View-consistent Generation

In the final stage, we adopt a view-consistent generation strategy based on a pre-trained Flow Transformer (Labs 2024), combining the coarse guidance (Sec. 3.2) and the iteratively trained LoRA module (Sec. 3.3) to produce the final consistent guidance images. As shown in Figure 3, we begin by rendering the coarse guidance into multi-view images and converting them into noisy latents using rectified-flow inversion (Rout et al. 2024) to encode both the appearance and geometry. Serving as the starting point for subsequent generation process, these latents are then fed into the Flow Transformer along with rendered depth maps, which provide geometric cues to align appearance generation with the underlying structure and enhance multi-view consistency.

Inspired by (Chen, Laina, and Vedaldi 2024; Feng et al. 2025), we introduce an epipolar-constrained token replacement mechanism to promote multi-view consistency by unifying foreground tokens across views that correspond to the same 3D locations. Specifically, we perform token replacement in the early dual-stream blocks of the Flow Transformer, where visual tokens are explicitly maintained and can be directly modified. The mechanism starts by automatically selecting a set of key views with minimal overlap to ensure full coverage, and extracting all foreground pixel indices using multi-view masks from the coarse guidance. As illustrated in Figure 4(b), for each foreground token in non-key views, we compute its epipolar lines on the two nearest key views and replace it with the interpolation of its most similar tokens along the two epipolar lines, weighted by its camera distance to the key views. Given a non-key frame, the interpolation $\mathbf{f}'(\mathbf{u})$ used to replace the original token $\mathbf{f}(\mathbf{u})$ at pixel $\mathbf{u}$ is computed using the foreground tokens in the two nearest key frames $\mathbf{k}_i$, indexed by $i \in \{1, 2\}$. Letting c denote the corresponding camera and $l_{\mathbf{u}\to i}$ denote the epipolar line of $\mathbf{u}$ in the i-th key view, the token $\mathbf{f}'(\mathbf{u})$ is computed as:

$$\mathbf{f}'(\mathbf{u}) = \sum_i \mathbf{k}_i(\mathbf{v}_i)\mathcal{D}(c, c_i) / \sum_i \mathcal{D}(c, c_i),$$
$$\text{where } \mathbf{v}_i = \arg\max_{\mathbf{v}\in l_{\mathbf{u}\to i}} \langle \mathbf{f}(\mathbf{u}), \mathbf{k}_i(\mathbf{v}) \rangle \quad (3)$$

where $\mathcal{D}(c, c_i)$ represents the camera distance from c to the i-th key view's camera c_i. This mechanism effectively alleviates cross-view variance, producing fine-grained guidance images with strong multi-view consistency and faithful reference alignment. These images then supervise the 3DGS parameter updating of the coarse guidance, yielding the final personalized 3DGS result of our CP-GS framework.

4 Experiments

4.1 Implementation Details

We implement our framework based on the official 3DGS codebase (Kerbl et al. 2023), GaussianEditor (Chen et al. 2023b), and the LoRA training scripts from Diffusers (von Platen et al. 2022). We employs TRELLIS (Xiang et al. 2024) as image-to-3D model to generate our coarse guidance. For the Flow transformer, we adopt FLUX.1-dev (Labs 2024) equipped with the depth LoRA adapter. In most cases, we use two iterations of LoRA training, which takes around 15 minutes on two NVIDIA RTX A6000 GPUs and is reusable across different source scenes. Using the trained LoRA and coarse guidance, we generate consistent multi-view guidance images and optimize the 3DGS model with the Adam optimizer at a learning rate of 0.001, taking around 10 minutes per scene when using the same two A6000 GPUs.

4.2 Qualitative Evaluation

We compare our CP-GS with two state-of-the-art 3DGS editing baselines: DGE (Chen, Laina, and Vedaldi 2024), which conditioned on text prompt, and TIP-Editor (Zhuang et al. 2024), the only existing method with the same single-image condition as ours. We construct a challenging test set comprising reference images from TIP-Editor and additional internet-sourced examples with highly specialized, visually intricate appearances. For DGE, we employ GPT-4o (Hurst et al. 2024) to generate concise captions (within 5 words) that describe the reference object, as longer prompts were observed to degrade its performance. As shown in Figure 5, both baselines fail to preserve the distinctive appearance features of the reference images. Moreover, TIP-Editor exhibits severe artifacts in its personalized results, primarily due to multi-view inconsistencies in the guidance images resulting from viewpoint bias. In contrast, our CP-GS consistently produces clean, coherent, and intricately detailed edits that faithfully align with the reference image.

This performance gap underscores the baseline methods' inability to capture and propagate reference appearance. DGE

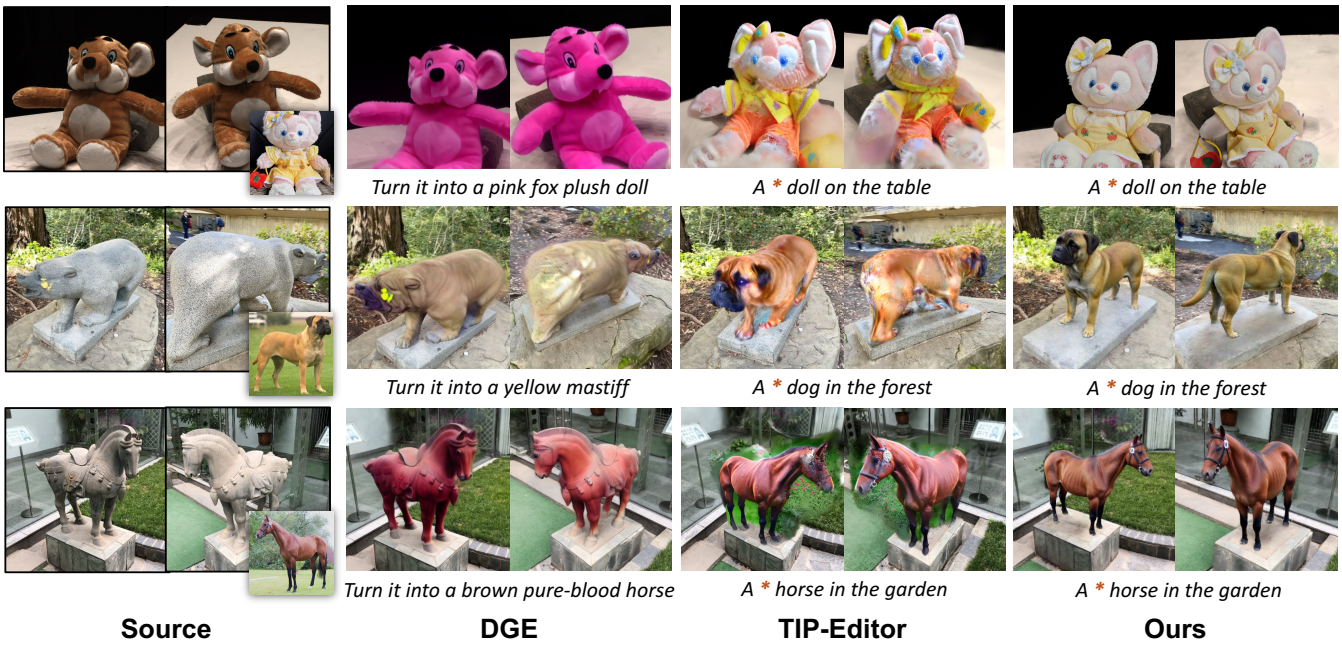

Figure 5: Qualitative comparison of personalization results between our CP-GS and the existing SOTA methods (Chen, Laina, and Vedaldi 2024; Zhuang et al. 2024), where CP-GS outperforms with superior visual quality and reference alignment.

illustrates the shortcomings of text-conditioned 3DGS editing for personalization: lacking direct access to the reference image, it relies solely on short textual prompts that fail to capture rich visual details. Moreover, the specialized reference images often fall outside the distribution of text-to-image models, making them difficult to represent accurately. On the other hands, TIP-Editor lacks explicit mechanism to extend the reference appearance to novel viewpoints, resulting in strong viewpoint bias, which introduces multi-view inconsistencies in its 2D guidance, ultimately leading to visual artifacts in the 3DGS results. By contrast, CP-GS explicitly addresses these issues through the coarse-to-fine appearance propagation, enabling high-quality 3DGS personalization that ensures both referential and multi-view consistency. Additional results showcasing the effectiveness of our CP-GS are presented in the *Appendix*.

4.3 Quantitative Evaluation

In Table 1, we present a quantitative evaluation comparing our CP-GS with the two baselines (Chen, Laina, and Vedaldi 2024; Zhuang et al. 2024) on over 20 samples collected in the same manner as the qualitative experiments. We first conduct a user study to assess the proportion of results deemed satisfactory by users in aspects of visual quality and the reference alignment, with further details provided in the *Appendix*. Besides, we follow existing setting (Zhuang et al. 2024) to report CLIP (Radford et al. 2021) and DINO (Oquab et al. 2023) image-to-image similarity metrics that quantify the alignment between edited outputs and the reference image by computing visual feature similarity. In addition, we adopt the CLIP directional similarity (Gal et al. 2022b) to measure the semantic alignment between the text prompt and the semantic shift from the source to edited results. As shown in Table 1, our CP-GS consistently outperforms all baselines across both perceptual metrics and user study evaluations. This performance gap stems from two main limitations in these baselines: (1) DGE relies solely on text modality as condition, which fails to capture fine-grained appearance details; (2) TIP-Editor fails to propagate the reference appearance to novel views, resulting in strong viewpoint bias that introduces multi-view conflicts in the editing guidance and ultimately leads to visual artifacts and poor reference alignment. These results highlight the superior performance of our CP-GS, underscoring the critical role of capturing and expanding reference appearance across novel viewpoints in tackling single-view conditioned 3DGS personalization.

4.4 Ablation Study

To analysis the contribution of each component, in this section, we compare the coarse guidance with our final results and conduct ablation studies on the iterative LoRA fine-tuning strategy and epipolar-constrained token replacement. An additional quantitative ablation study is in *Appendix*.

Coarse Guidance vs. Final Results. In the first stage of our method, a coarse asset is generated by the image-to-3D model (Sec. 3.2) and integrated into the source scene to produce the coarse guidance. A natural question arises: *can this coarse guidance suffice as the final edit?* To assess the necessity of our subsequent stages, we compare the coarse guidance with our final personalization results in Figure 6 *left*. This comparison reveals two major limitations of the coarse guidance: (1) The domain gap between real-world reference images and the CGI-style training data of the 3D generation model (Xiang et al. 2024) results in overly smooth and grid-like unreal-

Figure 6: ***left***: Comparison between the coarse guidance and our final results, where our final results effectively refine the unrealistic and visually discordant appearance presented in the coarse guidance. ***right***: Ablation comparison on the guidance images produced by three specific configurations: training the LoRA module only on the single-view reference image (*Sing. LoRA*), using our iterative LoRA fine-tuning yet excluding the constrained token replacement (*Iter. LoRA*), and our *Full Version*.

Methods	$\text{User}_{quality}\uparrow$	$\text{User}_{align}\uparrow$	$\text{DINO}_{sim}\uparrow$	$\text{CLIP}_{sim}\uparrow$	$\text{CLIP}_{dir}\uparrow$
DGE (Chen, Laina, and Vedaldi 2024)	31.89%	6.37%	41.73	67.26	14.22
TIP-Editor (Zhuang et al. 2024)	25.46%	17.28%	43.88	70.92	14.46
CP-GS (Ours)	**78.28%**	**80.09%**	**50.33**	**76.78**	**18.03**

Table 1: Quantitative comparison between our CP-GS and the existing SOTA methods (Chen, Laina, and Vedaldi 2024; Zhuang et al. 2024), where CP-GS significantly outperform others in both visual quality and the alignment with reference image.

istic textures that fail to reflect the photorealic appearance of the reference. (2) Direct insertion of the generated asset leads to poor contextual blending, where the inserted object often appears visually detached from the 3DGS scene (e.g., the sunglasses example), especially around boundaries. In contrast, our final results exhibit rich, realistic textures that closely resemble the reference image and blend seamlessly with the source 3DGS scene, demonstrating the effectiveness and necessity of the subsequent refinement stages.

Iterative LoRA Fine-tuning. We compare two configurations to assess the contribution of our iterative LoRA fine-tuning: using only the single-view reference image for fine-tuning (*Sing. LoRA*) versus applying our iterative expansion strategy (*Iter. LoRA*) introduced in Sec. 3.3. The *top 2 rows* in Figure 6 *right* shows the resulting guidance images from these variants. In the presented viewpoints that deviate from the reference, *Sing. LoRA* misprojects the appearance entangled with reference-view geometry to unrelated views, resulting in noticeable distortion and multi-view inconsistency. In contrast, *Iter. LoRA* significantly alleviates this distortion, generating appearance correctly adapted to these novel perspectives. This highlights the importance of progressively expanding the reference coverage in mitigating the viewpoint bias and producing multi-view consistent guidance images.

Constrained Token Replacement. To further reduce the cross-view variance during generation, we adopt an epipolar-constrained token replacement strategy during the generation of final guidance images (Sec. 3.4). The *bottom 2 rows* in Figure 6 *right* compares the *Full Version* including this mechanism and the *Iter. LoRA* variant where it is disabled. Although *Iter. LoRA* successfully expands the reference appearance to novel viewpoints and resolves major distortions, it still suffers from subtle multi-view inconsistencies in visual details (e.g., the mouth orientation of the *dog*, and the *doll*'s eyelashes). In contrast, our *Full Version* leverages 3D-aware token replacement guided by epipolar constraints and eliminate such inconsistencies, producing guidance images with improved multi-view appearance consistency that enhance the coherence and visual fidelity of the final 3DGS results.

5 Conclusion

In this paper, we presented CP-GS, a novel framework for consistent and personalized 3D scene editing from a single-view reference image. To address the visual artifacts in existing image-conditioned methods caused by viewpoint bias and limited reference perspective, CP-GS introduces a coarse-to-fine reference propagation framework that integrates coarse guidance generation, iterative LoRA fine-tuning, and a view-consistent generation stage leveraging geometric cues and epipolar-constrained token replacement. These components enable the generation of guidance images with strong multi-view consistency and faithful referential consistency, producing high-quality 3DGS results. Extensive experiments across real-world scenes demonstrate that CP-GS markedly outperforms existing methods in both visual quality and reference alignment, enabling high-quality 3DGS personalization for a variety of real-world applications in 3D industry.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund — Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).

References

Abdi, H.; and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4): 433–459.

Avrahami, O.; Aberman, K.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, 1–12.

Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; and Kim, S. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3558–3568.

Chen, H.; Zhang, Y.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2023a. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 3(4).

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Dge: Direct gaussian 3d editing by consistent multi-view editing. In *European Conference on Computer Vision*, 74–92. Springer.

Chen, X.; Feng, Y.; Chen, M.; Wang, Y.; Zhang, S.; Liu, Y.; Shen, Y.; and Zhao, H. 2024a. Zero-shot image editing with reference imitation. *Advances in Neural Information Processing Systems*, 37: 84010–84032.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024b. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6593–6602.

Chen, Y.; Chen, Z.; Zhang, C.; Wang, F.; Yang, X.; Wang, Y.; Cai, Z.; Yang, L.; Liu, H.; and Lin, G. 2023b. GaussianEditor: Swift and Controllable 3D Editing with Gaussian Splatting. arXiv:2311.14521.

Chen, Z.; Fang, S.; Liu, W.; He, Q.; Huang, M.; and Mao, Z. 2024c. DreamIdentity: enhanced editability for efficient face-identity preserved image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38.

Choi, J.; Choi, Y.; Kim, Y.; Kim, J.; and Yoon, S. 2023. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; et al. 2023a. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023b. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13142–13153.

Dihlmann, J.-N.; Engelhardt, A.; and Lensch, H. 2024. Signerf: Scene integrated generation for neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6679–6688.

Dong, J.; and Wang, Y.-X. 2024. ViCA-NeRF: View-Consistency-Aware 3D Editing of Neural Radiance Fields. arXiv:2402.00864.

Edstedt, J.; Sun, Q.; Bökman, G.; Wadenbäck, M.; and Felsberg, M. 2024. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19790–19800.

Fang, J.; Wang, J.; Zhang, X.; Xie, L.; and Tian, Q. 2023. GaussianEditor: Editing 3D Gaussians Delicately with Text Instructions. arXiv:2311.16037.

Feng, H.; Huang, Z.; Li, L.; Lv, H.; and Sheng, L. 2025. Personalize anything for free with diffusion transformer. *arXiv preprint arXiv:2503.12590*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022a. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022b. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Haque, A.; Tancik, M.; Efros, A. A.; Holynski, A.; and Kanazawa, A. 2023. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. arXiv:2303.12789.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kamata, H.; Sakuma, Y.; Hayakawa, A.; Ishii, M.; and Narihira, T. 2023. Instruct 3D-to-3D: Text Instruction Guided 3D-to-3D conversion. *arXiv preprint arXiv:2303.15780*.

Karim, N.; Khalid, U.; Iqbal, H.; Hua, J.; and Chen, C. 2023. Free-Editor: Zero-shot Text-driven 3D Scene Editing. *arXiv preprint arXiv:2312.13663*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Koo, J.; Park, C.; and Sung, M. 2023. Posterior Distillation Sampling. arXiv:2311.13831.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Lee, D. I.; Park, H.; Seo, J.; Park, E.; Park, H.; Baek, H. D.; Sangheon, S.; Kim, S.; and Kim, S. 2024. EditSplat: Multi-View Fusion and Attention-Guided Optimization for View-Consistent 3D Scene Editing with 3D Gaussian Splatting. *arXiv preprint arXiv:2412.11520*.

Li, D.; Li, J.; and Hoi, S. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36: 30146–30166.

Li, T.; Ku, M.; Wei, C.; and Chen, W. 2023. Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*.

Liu, Z.; Feng, R.; Zhu, K.; Zhang, Y.; Zheng, K.; Liu, Y.; Zhao, D.; Zhou, J.; and Cao, Y. 2023. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Park, J.; Kwon, G.; and Ye, J. C. 2023. ED-NeRF: Efficient Text-Guided Editing of 3D Scene using Latent Space NeRF. *arXiv preprint arXiv:2310.02712*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Peng, X.; Zhu, J.; Jiang, B.; Tai, Y.; Luo, D.; Zhang, J.; Lin, W.; Jin, T.; Wang, C.; and Ji, R. 2024. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27080–27090.

Pexels. 2024. The best free stock photos, royalty free images & videos shared by creators. https://www.pexels.com.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raj, A.; Kaza, S.; Poole, B.; Niemeyer, M.; Ruiz, N.; Mildenhall, B.; Zada, S.; Aberman, K.; Rubinstein, M.; Barron, J.; et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2349–2359.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rout, L.; Chen, Y.; Ruiz, N.; Caramanis, C.; Shakkottai, S.; and Chu, W.-S. 2024. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Sohn, K.; Ruiz, N.; Lee, K.; Chin, D. C.; Blok, I.; Chang, H.; Barber, J.; Jiang, L.; Entis, G.; Li, Y.; et al. 2023. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*.

Song, L.; Cao, L.; Gu, J.; Jiang, Y.; Yuan, J.; and Tang, H. 2023. Efficient-NeRF2NeRF: streamlining text-driven 3D editing with multiview correspondence-enhanced diffusion models. *arXiv preprint arXiv:2312.08563*.

von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; Nair, D.; Paul, S.; Berman, W.; Xu, Y.; Liu, S.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Wang, Y.; Yi, X.; Wu, Z.; Zhao, N.; Chen, L.; and Zhang, H. 2024. View-consistent 3d editing with gaussian splatting. In *European Conference on Computer Vision*, 404–420. Springer.

Wu, J.; Bian, J.-W.; Li, X.; Wang, G.; Reid, I.; Torr, P.; and Prisacariu, V. A. 2024. Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing. In *European Conference on Computer Vision*, 55–71. Springer.

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. *arXiv preprint arXiv:2412.01506*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18381–18391.

Yi, X.; Wu, Z.; Xu, Q.; Zhou, P.; Lim, J.-H.; and Zhang, H. 2024. Diffusion time-step curriculum for one image to 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9948–9958.

Yuan, Z.; Cao, M.; Wang, X.; Qi, Z.; Yuan, C.; and Shan, Y. 2023. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint arXiv:2310.19784*.

Zhuang, J.; Kang, D.; Cao, Y.-P.; Li, G.; Lin, L.; and Shan, Y. 2024. TIP-Editor: An Accurate 3D Editor Following Both Text-Prompts And Image-Prompts. *arXiv preprint arXiv:2401.14828*.

Zhuang, J.; Wang, C.; Liu, L.; Lin, L.; and Li, G. 2023. DreamEditor: Text-Driven 3D Scene Editing with Neural Fields. arXiv:2306.13455.