In-N-Out: Lifting 2D Diffusion Prior for 3D Object Removal via Tuning-Free Latents Alignment

Dongting Hu [1] **Huan Fu** [3] **Jiaxian Guo** [4] **Liuhua Peng** [1]
Tingjin Chu [1] **Feng Liu** [1] **Tongliang Liu** [2, 5] **Mingming Gong** [1, 5]

[1] The University of Melbourne [2] The University of Sydney [3] Alibaba
[4] Google Research [5] Mohamed bin Zayed University of Artificial Intelligence
Project Page: https://timmy11hu.github.io/3dor.github.io/

Abstract

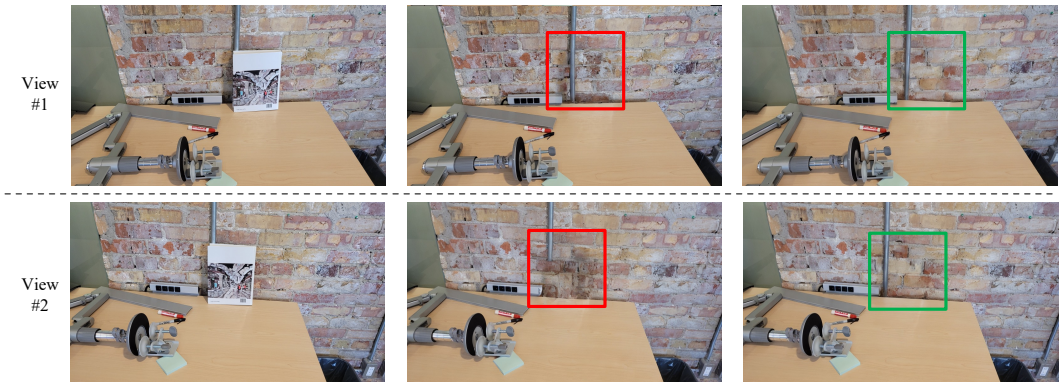
Neural representations for 3D scenes have made substantial advancements recently, yet object removal remains a challenging yet practical issue, due to the absence of multi-view supervision over occluded areas. Diffusion Models (DMs), trained on extensive 2D images, show diverse and high-fidelity generative capabilities in the 2D domain. However, due to not being specifically trained on 3D data, their application to multi-view data often exacerbates inconsistency, hence impacting the overall quality of the 3D output. To address these issues, we introduce "In-N-Out", a novel approach that begins by inpainting a prior, i.e., the occluded area from a single view using DMs, followed by outstretching it to create multi-view inpaintings via latents alignments. Our analysis identifies that the variability in DMs' outputs mainly arises from initially sampled latents and intermediate latents predicted in the denoising process. We explicitly align of **initial** latents using a Neural Radiance Field (NeRF) to establish a consistent foundational structure in the inpainted area, complemented by an implicit alignment of **intermediate** latents through cross-view attention during the denoising phases, enhancing appearance consistency across views. To further enhance rendering results, we apply a patch-based hybrid loss to optimize NeRF. We demonstrate that our techniques effectively mitigate the challenges posed by inconsistencies in DMs and substantially improve the fidelity and coherence of inpainted 3D representations.

1 Introduction

Neural Radiance Fields (NeRFs) [50, 2, 23, 58, 81, 37, 13, 10, 3, 92] have effectively revolutionized 3D scene reconstruction from multi-view images. These models offer high-fidelity novel-view synthesis, proving beneficial across a variety of domains [32, 87, 88, 43, 107, 6, 63, 72, 8, 60]. Despite the impressive ability to reconstruct highly detailed scenes, these learning-based methods depend on the availability of consistent multi-view training data. This reliance limits their generalizability, particularly in editing 3D representations for tasks like object removal and inpainting occluded areas.

Recently, diffusion models (DMs) [30, 18, 71, 80] have gained significant attention in the field of generative modelling for 2D images. These models are well-known for their robustness as generative priors, capable of producing diverse and high-fidelity results in 2D inpainting tasks. However, adapting these 2D priors for 3D object removal is not straightforward. While the inherent diversity of DMs benefits the generation of varied outputs, it also poses a significant challenge: high variance in the inpainted results (Fig.1 middle column). Consequently, these models frequently produce outputs that, while visually appealing in isolation, may appear misaligned when incorporated into 3D domain [53, 94, 93, 25, 19, 97, 84]. This misalignment often results in the loss of high-frequency details, crucial for realistic and coherent scene rendering.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) Original Image (b) Inpainted by Stable Diffusion (c) Inpainted by Ours

Figure 1: Inpainting outcomes of multi-view images from original Stable Diffuion [71] (middle) with those achieved by our approach (right). The inpainted areas are highlighted in red and green boxes.

Previous studies addressing such 3D inconsistencies can be broadly categorized into two approaches: multi-view and single-view priors. The former tackled inconsistencies across multi-view inpainted images by optimizing NeRFs with modified objectives [53, 94, 93, 65, 91]. While these methods have shown promise in refining inconsistent inputs, they sometimes suffer from a loss of detail fidelity during the training process, as illustrated in Fig. 4. Conversely, other studies have attempted to overcome the multi-view inconsistency bottleneck by anchoring the inpainting process to a single reference image that serves the entire scene [42, 51, 109]. This approach, however, places significant reliance on the selection of an appropriate reference image and the accuracy of depth estimates, which could lead to geometric artifacts during testing, as shown in Fig. 4.

To address these challenges, we aim to overcome 3D inconsistencies by guiding 2D DMs to achieve multi-view consistent inpainting results (Fig. 1 right column). Our analysis reveals that the variance in model outputs primarily comes from the random noise as the **initial** latent sample, and **intermediate** latents inferred by the denoising network. Each frame's initial latents are independently sampled, while intermediate latents are individually predicted, highlighting how view-dependent data impacts the generation process. Therefore, our approach focuses on aligning these two critical elements across multiple inputs. We introduce "In-N-Out", a conditional-sampling-like approach that inpaints a sampled view and outstretches it to multiple views. Our method contains three key components:

1. Conditional Inpainting Pipeline: We propose a pipeline that first samples an inpainting outcome from a random view as an inpainting prior. This prior then serves as a condition to guide the inpainting process of multiple views, ensuring a consistent inpainting foundation.

2. Explicit Latents Alignment: Leveraging the geometry derived from a pre-trained NeRF and the inpainting prior, we sample multi-view initial latents conditional on the geometry dictated by the inpainting prior. This ensures that the primary components within the inpainted areas are structurally consistent and align with the underlying 3D geometry.

3. Implicit Latents Alignment: We employ a cross-view attention mechanism during the denoising steps to align predicted intermediate latents concerning the inpainting prior. This enhances the appearance consistency across the inpainted images.

To further enhance our method's performance in the 3D domain, we have implemented a patch-based optimization strategy using a hybrid loss on our inpainted multi-view images. This strategy employs perceptual loss to rectify spatial mismatches, and adversarial loss to preserve high-frequency details. By addressing these key challenges, our framework effectively handles multi-view inconsistencies and enhances the fidelity and coherence of 3D representations. The effectiveness of our approach is demonstrated through both qualitative and quantitative evaluations of a challenging object removal dataset. Our results indicate comprehensive improvement compared to existing methods, highlighting our model's ability to achieve greater fidelity and consistency in inpainted scenes.

2 Related Works

2D Editing with Diffusion Models Diffusion models [30, 33, 105, 59, 79, 80], have revolutionized image generation with their capacity to create highly realistic images. These models facilitate customizable generation via textual prompts [18, 29, 69, 74], predominantly using pre-trained Stable Diffusion [71]. Several editing methods [22, 57, 56, 61] allow users to adjust images by moving anchor points to new locations. Editing typically begins by inverting the latent representation of the image to be edited back to its initial noise [80], with modifications made during the denoising phase. Prompt-to-Prompt (P2P) [26] edits images by adjusting the cross-attention between the image and text. Null-text inversion [55] addresses artifacts in DDIM inversion [80] when using classifier-free guidance [29]. Delta Denoising Score (DDS) [27] optimizes the latent image representation by aligning the predicted noises of the original and modified texts. Additionally, several studies [7, 9, 26] have identified a relationship between the appearance of images generated by diffusion models and the key-value pairs. While these advancements represent significant progress preserve some content from the original image in 2D image editing, they do not account for multi-view consistency, thus can not be lifted to 3D editing directly.

Lifting 2D diffusion models for 3D editing Recent advancements in 3D editing and generation have effectively utilized 2D DMs to enhance these processes, as demonstrated in various studies [52, 76, 96, 62, 97, 99, 41, 95, 68, 49, 103]. Pioneering works have used images inferred by DMs for direct supervision. Instruct-NeRF2NeRF (IN2N) [25] approached the editing task by transforming 3D model editing into a 2D image editing task, utilizing Instruct Pix2Pix (IP2P) [5] to iteratively update 3D scenes. Similarly, ViCA-NeRF [19] addressed editing challenges by modifying reference images and integrating these changes into the scene. DreamEditor [110] opted for a different strategy by converting NeRF into a mesh for direct optimization. GaussianEditor [12] applies semantic tracing to identify and modify editing targets within 3D Gaussian Splatting (3DGS) [37]. Similarly, Gaussian Grouping [100] implements Identity Encoding for each Gaussian to create masks for editing. Conversely, Score Distillation Sampling (SDS) [64] provides an alternative way to guide 3D representations by backpropagating gradients from a diffusion model's denoiser [1, 16, 73, 71] into the underlying scene representation. This technique has been effectively applied to generate realistic 3D and 4D scenes using NeRFs [11, 40, 39, 108, 110] and 3DGS [70, 101, 84, 15].

2D and 3D Inpainting 2D inpainting methods reconstruct images by filling missing content in areas defined by a mask [20, 77, 104, 86, 46, 75, 102]. Early techniques, exemplified by [21], relied on copying textures from known to unknown regions. LaMa [82] excels in restoring large missing areas using fast Fourier convolutions, extensive receptive fields, and large training masks. Although highly effective at generating plausible background textures within specified masks, LaMa limits the fidelity of its outputs. In contrast, probabilistic diffusion models [30] have shown impressive results in image generation and offer a wide range of inpainted outputs. DMs can be adapted for inpainting without specific training, and modify known regions during each denoising step to fit the task [47]. Similarly, Stable Diffusion [71] excels at inpainting by operating within latent space, allowing for efficient and effective image generation. In this work, we adopt it as our 2D inpainter.

3D scene inpainting aims to fill missing areas within a space, such as removing objects and generating coherent geometry and textures to complete the scene. Although 3D generative models have garnered large interest [4, 34, 38, 89, 78, 44, 31, 85, 45, 14, 12], they are often limited by the scarcity of 3D training data, hence result in poor generalization, particularly in scene inpainting tasks. Therefore, most current 3D inpainting models [53, 51, 42, 65, 93, 94, 109, 91] enhance their effectiveness by adopting priors from 2D models. SPIn-NeRF [53] reduces multi-view inconsistencies by first inpainting views and then optimizing NeRF using perceptual loss. NeRFiller [93] tackles multiple frames simultaneously by tiling images for DMs. GaussianEditor [12] edit targets within 3DGS [37], guided by inpainted multi-view images from DMs. While these methods show promise, they can sometimes compromise detail fidelity during training. Alternatively, some studies circumvent multi-view inconsistencies by using a single reference image for the entire scene [42, 51, 109]. Infusion [109] stands out in the inpainting of 3DGS, leveraging a pre-trained depth completion network to infer point clouds from a single inpainted view, though this method depends heavily on precise depth estimates. Concurrent works [65, 54] address these challenges using SDS objective [64] to better align 2D model priors with 3D scene consistency.

3 Preliminaries

3.1 Neural Radiance Fields

Neural Radiance Fields (NeRFs) [50] represents a breakthrough in 3D rendering by employing a multilayer perceptron (MLP), denoted as ϕ to represent a scene. This MLP serves as a continuous volumetric function to capture and reconstruct a scene in unprecedented detail. Specifically, NeRFs take as input the view direction d and a 3D coordinate $r(\tau)$ sampled from a camera ray defined by $r(\tau) = o + \tau d$. At each position along this ray $r(\tau)$, the network predicts the volume density and view-dependent color, represented as (σ, c). To render a camera pixel, NeRFs perform an aggregation of the predicted densities and color emissions $\sigma(\tau_i), c(\tau_i)$ along the camera ray. This process is mathematically formulated as an approximation of a volume rendering integral [48], which is used to compute the final color of the pixel:

$$\hat{C}(r) = \sum \Gamma_i \left(1 - \exp\left(-\sigma(\tau_i)\delta(\tau_i)\right)\right) c(\tau_i), \quad \text{with } \Gamma_i = \exp\left(-\sum_{j=1}^{i-1} \sigma(\tau_j)\delta_j\right), \tag{1}$$

where $\delta(\tau_i) = \tau_{i+1} - \tau_i$ is the distance between adjacent samples along the ray. During the training phase, rays are uniformly sampled from the training images, and the volumetric field is optimized using mean square error (MSE) to enhance the accuracy and realism of the rendered scenes.

3.2 Diffusion Models

Diffusion models [30] consist of two processes: a forward process that gradually introduces noise to a data sample $z^0 \sim p_{\text{data}}(z)$, and a learned reverse process that iteratively denoises a purely Gaussian noise sample $z^T \sim \mathcal{N}(0, 1)$ back into a clean image z^0. The reverse process is parameterized by a conditional noise prediction network ϵ_θ, trained to predict the noise using the simplified objective:

$$p_\theta(z^{0:T}|c) = p(z^T) \prod_{t=1}^{T} p_\theta(z^{t-1}|z^t, c), \quad p_\theta(z^{t-1}|z^t, c) = \mathcal{N}(z^{t-1}; \mu_\theta(z^t, t, c), \sigma^2 I), \tag{2}$$

where t is the time step in the diffusion process, z^t is an intermediate noisy sample, and c represents a condition (e.g., images, masks, or text). Utilizing a deterministic sampler like DDIM [80], the sample z^{t-1} can be obtained by $z^{t-1} = z^t - \epsilon_\theta(z^t, t, c)$; note that scaling is omitted for simplicity. In practice, as we use Stable Diffusion [71], a latent diffusion as the inpainting backbone, z is latent and the generated image is obtained with a decoder $\Omega(z^0)$. Hence, the variability of the generated image z^0 depends solely on initial **sampled** latent z^T and intermediate **inferred** latents $\left\{z^{t-1}\right\}_{t=1}^{T}$.

4 Method

Given a set of multi-view training images $\{\mathcal{I}_i\}_{i=1}^N$ from the scene with corresponding masks $\{\mathcal{M}_i\}_{i=1}^N$ indicate the unwanted object in each frame, our approach seeks to generate consistently inpainted training set $\{\tilde{\mathcal{I}}_i\}_{i=1}^N$ and use them to supervise NeRF. Our approach is structured into three key stages:

- Stage 1: Pretrain a NeRF ϕ using $\{\mathcal{I}_i\}_{i=1}^N$ and $\{\mathcal{M}_i\}_{i=1}^N$, along with a sampled inpainted prior $\tilde{\mathcal{I}}_p$ as a rough hallucination of the inpaint feature. (Sec. 4.1).
- Stage 2: Leverage ϕ to inpaint additional views $\{\tilde{\mathcal{I}}_i \mid i \neq p, i = 1, \ldots, N\}$ conditioned on the inpainting prior $\tilde{\mathcal{I}}_p$ via explicit and implicit latents alignment. (Sec. 4.2)
- Stage 3: Using the inpainted image set $\{\tilde{\mathcal{I}}_i\}_{i=1}^N$, we optimize ϕ with a patch-based hybrid loss to distill multi-view supervision. (Sec. 4.3)

An overview of our method is shown in Fig. 2.

4.1 Stage 1: Pre-train NeRF

The initial stage involves training the NeRF on the unmasked region, we follow the original work [50] where simple MSE loss is applied:

$$\mathcal{L}_{\text{rec}}(\phi) = \sum_{r \in R_{\text{unmasked}}} \left\| \hat{C}_\phi(r) - C(r) \right\|_2^2, \tag{3}$$

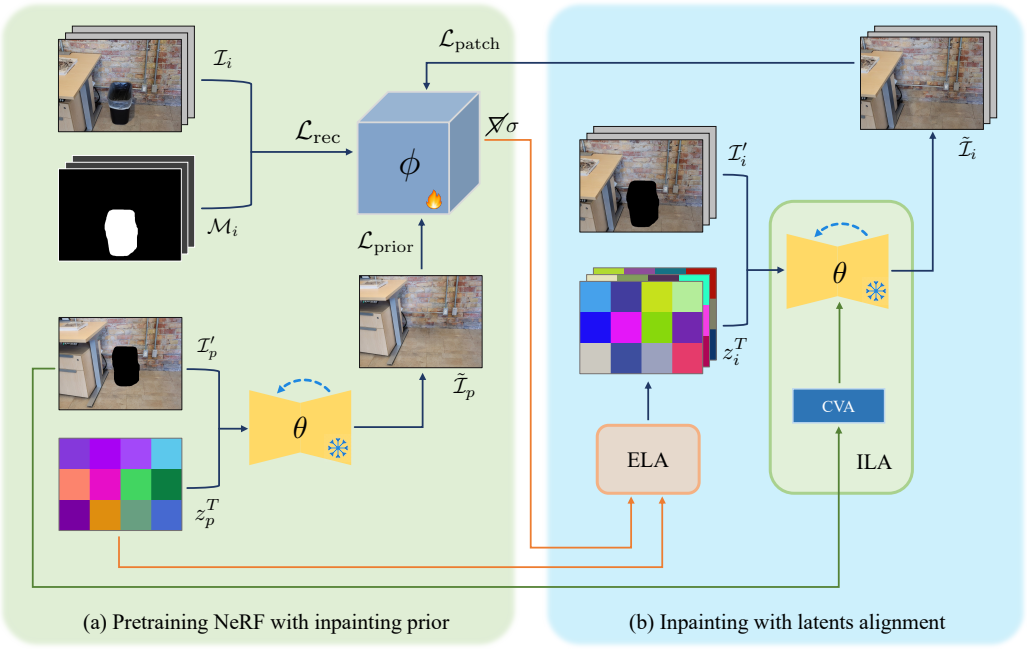


Figure 2: Overview of our method. Our approach begins with (a) pre-training the NeRF ϕ with a sampled inpainting prior $\tilde{\mathcal{I}}_p$ from Stable Diffusion θ, detailed in Sec. 4.1. It then progresses to (b) latent-aligned inpainting $\tilde{\mathcal{I}}_i$ for multi-view images through Explicit Latents Alignment (ELA) and Implicit Latents Alignment (ILA), as described in Sec. 4.2. Finally, the NeRF is optimized using a patch-based hybrid loss strategy outlined in Sec. 4.3. Throughout the training process, we fix Stable Diffusion θ and update the scene-specific NeRF parameters ϕ only.

where R_{unmasked} represent the unmasked pixels across all the training images. Then we sample a prior view $\mathcal{I}_p$ with its mask $\mathcal{M}_p$ and regularly inpaint it using Stable Diffusion θ. For illustration, we replace of condition in Eq. 2 with two components used by Stable Inpainting Diffusion [71] as e for the input prompt, and $\mathcal{I}'_p$ for the masked image that fed into the diffusion models. Hence the inpainting process can be formulated as:

$$z_p^{t-1} = z_p^t - \epsilon_\theta(z_p^t, t, \mathcal{I}'_p, e), \quad \text{for } t = T, \dots, 1, \quad \text{with } z_p^T \sim \mathcal{N}(0, 1). \tag{4}$$

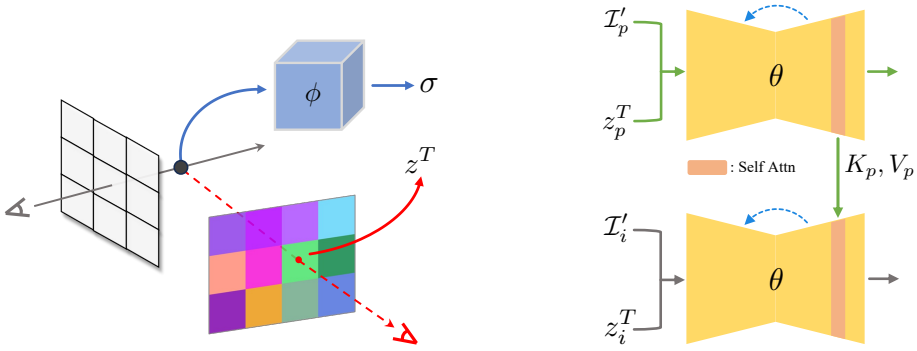
The inpainted image then can be obtained by $\tilde{\mathcal{I}}_p = \Omega(z_p^0)$. We then use a monocular depth estimator on $\tilde{\mathcal{I}}_p$ to get a depth map $\tilde{D}_p$. We regress the scale and offset parameters to align $\tilde{D}_p$ with the depth estimated from field ϕ on the unmasked pixels. Hence, we can introduce the geometry and appearance supervision of the inpainting prior $\tilde{\mathcal{I}}_p$ into the NeRF's optimization through:

$$\mathcal{L}_{\text{prior}}(\phi) = \sum_{r \in R_{\text{masked(p)}}} \left\| \hat{C}_\phi(r) - C(r) \right\|_2^2 + \left\| \hat{D}_\phi(r) - \tilde{D}_p(r) \right\|_2^2, \tag{5}$$

where $R_{\text{masked(p)}}$ denoted the masked (inpainted) pixels of $\tilde{\mathcal{I}}_p$, and $\hat{D}_\phi$ is the depth estimated by NeRF. This stage is depicted in Fig. 2(a).

4.2 Stage 2: Latents Alignment

In this section, we introduce our key approach to condition the additional inpainted frames to have an inpainting feature based on the prior $\tilde{\mathcal{I}}_p$. As discussed before, in deterministic sampling of the diffusion inpainting model θ, the generation structure and layout highly depend on (1) initial **sampled** noise z_i^T and (2) intermediate **predicted** latents $\{z_i^{t-1}\}_{t=1}^T$. Hence if we can align the latents from different views with the prior one, the model is likely to generate multi-view consistent latent z_i^0, hence image $\tilde{\mathcal{I}}_i$. In this section, we discuss how to align two terms respectively.



(a) Explicit Initial Latent Alignment (ELA) (b) Implicit Intermediate Latents Alignment (ILA)

Figure 3: Illustration of two types of Latent Alignment. This figure depicts the Explicit Latents Alignment (ELA) and Implicit Latents Alignment (ILA) processes, as detailed in Sec. 4.2.

Explicit Initial Latent Alignment (ELA) Given the sampled latent of prior view z_p^T used in Stage 1 (Sec. 4.1), we could leverage geometric information to explicitly align the initial latent in 3D space. Given that estimated depth $\tilde{D}_p$, one possible solution is to warp the z_p^T into other views using camera matrices. However, $\tilde{D}_p$ is not guaranteed to be accurate hence such hard projection could yield significant errors. Alternatively, we propose to leverage the pre-trained NeRF ϕ, as it's a naturally 3D-consistent representation. Specifically, to sample a resolution-grain initial latent $z^T(r)$, we utilize the original formulation of volume rendering (Eq.1) but with the substitution of color c by latent z^T. We query the density σ from ϕ, and acquire z^T by reprojecting the sampled point to the image plane of the prior latent view z_p^T:

$$z^T(r) = \sum \Gamma_i \Big(1 - \exp\big(-\sigma(\tau_i)\delta(\tau_i)\big)\Big) z^T(\tau_i), \quad \text{with } z^T(\tau_i) = f_{p,i}(z_p^T, \tau_i), \tag{6}$$

where $f_{p,i}$ denote camera perspective projection according to p and i camera matrices. Such soft projection could avoid error accumulation in the inpainting process, and reduce the precision burden on the depth estimator. We illustrate this process in Fig. 3a. There are two key reasons why we propose fine-tuning the NeRF and using it as a geometric prior for ELA: (a) After finetuning the NeRF, the geometric is represented by NeRF as a sharp (low variance) unimodal distribution on the ray. Consequently, the aggregated feature remains sharp, preserving the variations in the initial latents. (b) We empirically found the depth prior inferred by the monocular depth estimator is not perfectly aligned with the NeRF. Fine-tuning the NeRF can also benefit this depth prior. Since NeRF learns relatively certain geometry in the known (unmasked) areas, this geometry constraint can improve the geometry of neighboring inpainted (masked) areas due to their geometric proximity. We compromise the view-dependent effect in NeRF within the ELA module. Due to the heuristic nature of diffusion models, incorporating such view-dependent effects into diffusion models' output remains elusive.

Implicit Intermediate Latents Alignment (ILA) While the initial latent could be aligned using the explicit method, intermediate latents are predicted by denoising network ϵ_θ which is hard to control. We address this issue by exploring the conditioning mechanism of the denoising network in Stable Diffusion [71]. Recall that in Eq. 4, denoising network ϵ_θ relies on the input prompt e and masked image $\mathcal{I}'_i$ to predict the noise occurrence in the current step. While we can use the unified prompt for all views to align the text condition in cross-attention of ϵ_θ, the masked images $\{\mathcal{I}'_i\}_{i=1}^N$ are inherently different due to multi-view nature. Note that $\mathcal{I}'_i$ condition is introduced based on spatial self-attention (SA) [90] in the U-Net:

$$\text{SA}(Q_i, K_i, V_i) = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i, \tag{7}$$

where Q_i obtained from each spatial resolution of the latent, K_i, V_i are derived from corresponding latent encoded from masked image $\mathcal{I}'_i$. We can impose the coherence of the denoising step by introducing cross-view attention (CVA) of the prior view (Fig. 3b):

$$\text{CVA}(Q_i, K_p, V_p) = \text{Softmax}\left(\frac{Q_i (K_p)^T}{\sqrt{d}}\right) V_p, \tag{8}$$

where K_p, V_p are from masked base image $\mathcal{I}'_p$. We can then implicitly align the denoising step by replacing the original SA with a weighted sum from SA and CVA, i.e. $\lambda_a * \text{SA}(Q_i, K_i, V_i) + (1 - \lambda_a) * \text{CVA}(Q_i, K_p, V_p)$. Through this technique we ensure that the intermediate latents $\{z_i^{t-1}\}_{t=1}^T$ are also conditioned on the prior $\{z_p^{t-1}\}_{t=1}^T$, while retain its distinctiveness due to the individual viewpoint. The rationale of replacing KV with "prior" p, but not Q is that the appearance information (V) of the prior image should be considered when inpainting the other views, with the amount of information propagation is weighted by its attention key value (K). The attention query value comes from the current inpainting view i, Q_i, representing the information the current inpainting for view i is searching for. Together with K_p, it decides how much attention the view i inpainter should place on the prior view, and finally incorporates the information of the prior view V_p into view i.

4.3 Stage 3: Joint Optimization

As the original intention of this work, we seek to distill the inpainted views into NeRF ϕ in a way such that the high-fidelity is preserved as much as possible as the unmasked region. While some priors work in 3D editing [25, 96, 97, 93] propose to update the training set iteratively until converge, we empirically find it not suitable for inpainting task since the loss of fidelity is significant and could fall into local optima. Hence we propose to inpaint a subset of training images at once and regard them as supplementary guidance using a patch-based hybrid loss:

$$\mathcal{L}_{\text{patch}}(\phi) = \sum_{\rho \in \mathcal{P}_{\text{sub}}} \left\| \hat{I}_\phi(\rho) - \tilde{\mathcal{I}}(\rho) \right\|_1 + \mathcal{L}_{\text{lpips}}(\hat{I}_\phi(\rho), \tilde{\mathcal{I}}(\rho)) + \mathcal{L}_{\text{adv}}(\hat{I}_\phi(\rho), \tilde{\mathcal{I}}(\rho)), \tag{9}$$

where ρ is a patch sample from the masked area of subset views $\mathcal{P}_{\text{sub}}$, $\hat{I}_\phi(\rho)$ is NeRF predicted patch, and $\mathcal{L}_{\text{lpips}}, \mathcal{L}_{\text{adv}}$ are perceptual distance LPIPS [106] and adversarial loss [24]. Here LPIPS is utilized to address geometry mismatches, while adversarial loss is employed to preserve high-frequency details. As shown in Fig. 2, the final optimization objective is:

$$\mathcal{L}(\phi) = \mathcal{L}_{\text{rec}}(\phi) + \mathcal{L}_{\text{prior}}(\phi) + \lambda_{\text{patch}} \mathcal{L}_{\text{patch}}(\phi). \tag{10}$$

The patches are uniformly sampled within the bounding box of the mask, with a size of 256×256. Therefore, only the inpainted area is being optimized by the patch loss.

5 Experiment

5.1 Evaluation Setting

Dataset: Aligning with methodologies employed in prior works [53, 51, 109], our experiments utilize the SPIn-NeRF dataset [53], selected for its comprehensive ground truth availability. This dataset is specifically designed for object removal evaluations and comprises 10 scenes. Each scene includes 60 images featuring an unwanted object (training views) and 40 images from which the object has been removed (test views). For both the training and test views, human-annotated masks indicating the object region are available. We further collected 9 forward-facing scenes with manually annotated masks to evaluate the effectiveness of our method. This dataset includes 4 indoor scenes and 5 outdoor scenes. In the training set, the masked region contains the unwanted object, while the test set contains the ground truth background in the masked region.

Baselines: In our study, we benchmark our method against a variety of established 3D inpainting approaches to ascertain its relative performance. These include the perceptual-based SPIn-NeRF [53], tiling-based NeRFiller [93] (both multi-view guidance) and InFusion [109] (single-view guidance). To ensure a fair comparison, we employ the same inpainting diffusion models across all methods and maintain consistency in the number of denoising steps and used prompts. We utilized the source code provided by the authors and ran all the methods using one NVIDIA A100 (80G) GPU.

Metrics: To quantitatively evaluate the effectiveness of our approach, we employ two similarity metrics: LPIPS [106] and FID [28]. Additionally, we use MUSIQ [36], a sharpness metric that quantifies the clarity and detail retention in the edited images. Following established protocols from previous studies [53], all metrics are calculated specifically within the bounding boxes defined by the masks, focusing the evaluation precisely on the regions most affected by the object removal task.

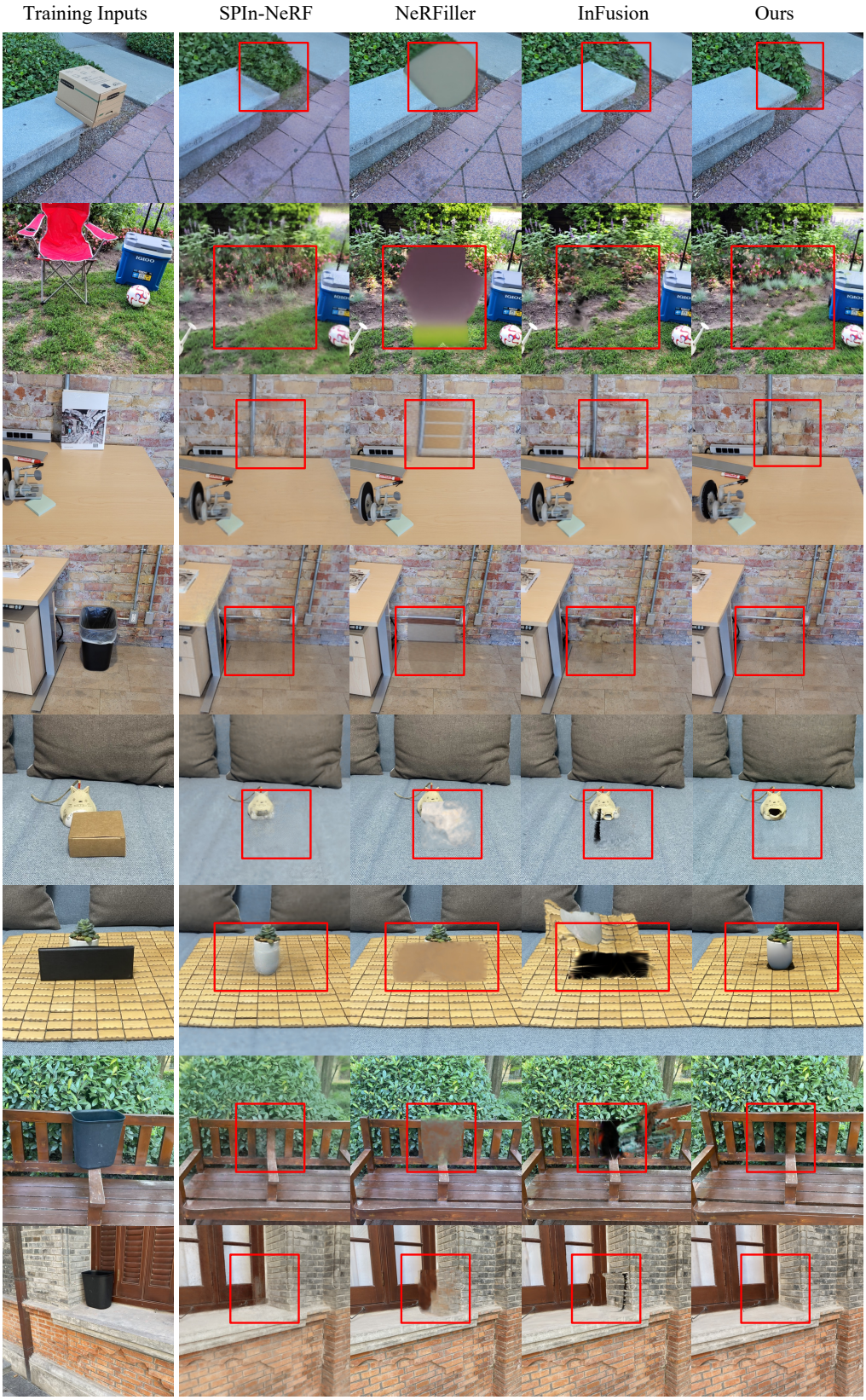


Figure 4: Qualitative results on the SPIn-NeRF the self-collected dataset.

5.2 Main Results

We first present the quantitative results in Tab. 1, where our method outperforms all baselines in terms of similarity metrics and sharpness. Our approach also excels in qualitative assessments, as demonstrated in Fig. 4. It is important to note that in simpler scenes with low variability in inpainting results, where the inconsistency issue is less pronounced (first row), most methods perform adequately. In other cases, high-frequency loss is observed in multi-view-based methods (SPIn-NeRF and NeRFiller). NeRFiller [93], through its use of multiple joint denoising steps, ensures consistency but often produces overly smooth outputs that lack fine details. It is noteworthy that the single-view-based method, InFusion [109], relies on one view and its depth to represent the entire scene. It performs well when geometry estimation is accurate. However, its performance deteriorates in scenarios where depth accuracy is compromised, leading to geometry artifacts (sixth and seventh rows). This underscores the critical role of multi-view supervision in addressing such challenges. By incorporating consistent multi-view supervision, our method remains effective even when depth or geometry is inaccurate, achieving robust and promising results. This explains why our method shows little difference from InFusion when the geometry is accurate (first and third row), but excels when the depth is inaccurate. Additionally, the exclusive reliance on perceptual loss by SPIn-NeRF [53] fails to fully address the multi-view inconsistencies introduced by inpainting diffusion models, often resulting in a blurred effect, particularly visible in the third and fourth rows. To further validate our findings, we conducted a user study based on the SPIn-NeRF dataset, focusing on the coherence of the background within the inpainted area, the fidelity of detail preservation in the inpainted region, and overall preference. The results of this study are summarized in Tab. 2. This evaluation clearly demonstrates superior performance across all assessed criteria.

Table 1: Quantitative Results

Method	LPIPS ↓	FID ↓	MUSIQ ↑
SPIn-NeRF [53]	0.54	185.63	38.69
NeRFiller [93]	0.71	315.83	32.60
InFusion [109]	0.62	153.77	39.29
Ours	**0.49**	**130.92**	**50.97**

Table 2: User Study

Method	Coherence	Fidelity	Overall
SPIn-NeRF [53]	22.72%	20.45%	21.82%
NeRFiller [93]	2.73%	4.33%	2.50%
InFusion [109]	27.50%	24.77%	25.00%
Ours	**47.05%**	**50.45%**	**50.68%**

5.3 Ablation Studies

We initially demonstrate the efficacy of our latents alignment approach with an example in Fig. 5. The first column displays the inpainting prior (sampled view), and the subsequent columns show the same training image being inpainted under different conditions. Notably, the variant without ELA (w/o ELA) retains colors similar to the prior but fails to preserve the texture structure. Conversely, the version without ILA (w/o ILA) maintains structural integrity but lacks appearance consistency with the prior. Our method effectively merges the strengths of both mechanisms, resulting in inpaintings that are highly consistent and cohesive across all evaluated aspects.

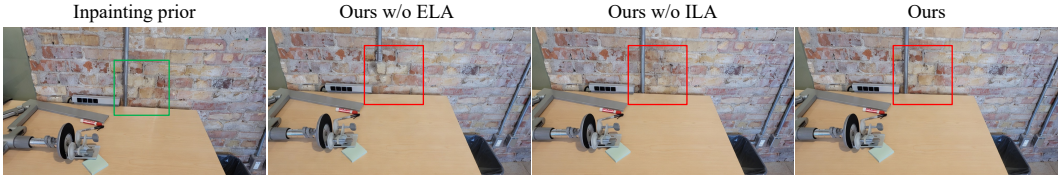


Figure 5: Ablation study on latent aligned onpainting. 2D Inpainting results when key components of our proposed method are omitted. Naive inpainting using Stable Diffusion can refer to Fig. 1.

We conducted further ablation studies to underscore the importance of our key design elements in object removal tasks. The quantitative and qualitative results, showcased in Tab. 3 and Fig. 6, clearly indicate the impact of each component. Notably, removing ELA leads to geometry mismatches in the NeRF outputs (w/o ELA), while deactivating ILA results in blurry coloration (w/o ILA). This observation confirms our initial findings: the initial latents primarily influence the inpainting's structural pattern, whereas the intermediate denoising steps largely affect its appearance, including

colour nuances. Additionally, our patch-based loss plays a crucial role in the optimization process (w/o $\mathcal{L}_{\text{patch}}$). Specifically, the $\mathcal{L}_{\text{lpips}}$ loss helps to alleviate geometry mismatches (w/o $\mathcal{L}_{\text{lpips}}$), and the $\mathcal{L}_{\text{adv}}$ serves as a detail-preserving supervisor (w/o $\mathcal{L}_{\text{adv}}$). These results highlight the effectiveness of our design choices in enhancing the overall quality and coherence of the inpainted outputs.

Table 3: Quantitative Results of Ablation Study.

Method	LPIPS ↓	FID ↓	MUSIQ ↑
Ours w/o ELA	0.52	133.09	48.90
Ours w/o ILA	0.50	141.78	49.70
Ours w/o $\mathcal{L}_{\text{patch}}$	0.73	293.32	33.76
Ours w/o $\mathcal{L}_{\text{lpips}}$	0.55	223.31	46.07
Ours w/o $\mathcal{L}_{\text{adv}}$	0.51	134.70	49.88
Ours full model	**0.49**	**130.92**	**50.97**

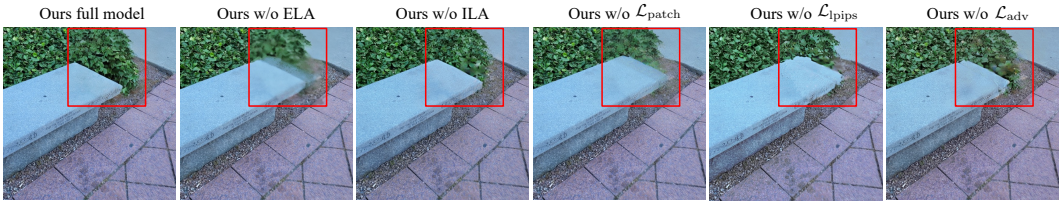


Figure 6: Ablation study on design choices based on rendering quality. This figure displays rendering results from NeRF when key components are individually removed from our full model.

6 Conclusion

In this work, we demonstrate the significant improvement achieved through our novel latents alignment approach in 3D object removal. By integrating both explicit and implicit latent alignment mechanisms, we have successfully addressed key challenges associated with geometry mismatches and color inconsistencies that are prevalent in the baselines, enhancing the fidelity and detail of the inpainted 3D scenes. The improvements achieved through our work offer significant societal benefits, such as enhanced editability of radiance fields. However, it also poses risks, including the potential perpetuation of biases and discrimination. If the data used to train diffusion models is biased, our approach could inadvertently reinforce these biases.

Despite notable advancements, our method has limitations: (1) It struggles with full 3D consistency, especially on high-frequency details, due to the constraints of applying 2D diffusion models to multi-view data. Future work could address this by integrating multi-view training into 2D inpainting diffusion models or leveraging true 3D generative models. (2) It is tailored for forward-facing scenes, limiting its applicability to diverse 360° views. Further exploration of latent relationships for broader view coverage is needed. (3) Predefined masks are currently required. Integrating advanced 3D perception methods [66, 67] could enhance accuracy and flexibility, enabling precise language-driven interactions and creating a more automated, user-friendly framework for neural 3D scene editing.

7 Acknowledgements

This research was mainly undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. This research was also partially supported by the Research Computing Services NCI Access scheme at the University of Melbourne. DH was supported by the Melbourne Research Scholarship from the University of Melbourne. FL is supported by the Australian Research Council (ARC) with grant numbers DP230101540 and DE240101089, and the NSF&CSIRO Responsible AI program with grant number 2303037.

References

[1] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[2] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. 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.

[3] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023.

[4] M. A. Bautista, P. Guo, S. Abnar, W. Talbott, A. Toshev, Z. Chen, L. Dinh, S. Zhai, H. Goh, D. Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems*, 35:25102–25116, 2022.

[5] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[6] A. Cao and J. Johnson. Hexplane: A fast representation for dynamic scenes. *CVPR*, 2023.

[7] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, October 2023.

[8] J. Cen, Z. Zhou, J. Fang, C. Yang, W. Shen, L. Xie, X. Zhang, and Q. Tian. Segment anything in 3d with nerfs. In *NeurIPS*, 2023.

[9] L. Cerkezi, A. Davtyan, S. Sameni, and P. Favaro. Multi-view unsupervised image generation with cross attention guidance. *arXiv preprint arXiv:2312.04337*, 2023.

[10] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022.

[11] R. Chen, Y. Chen, N. Jiao, and K. Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22246–22256, 2023.

[12] Y. Chen, Z. Chen, C. Zhang, F. Wang, X. Yang, Y. Wang, Z. Cai, L. Yang, H. Liu, and G. Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting, 2023.

[13] Z. Chen, T. Funkhouser, P. Hedman, and A. Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16569–16578, 2023.

[14] Y.-C. Cheng, H.-Y. Lee, S. Tulyakov, A. G. Schwing, and L.-Y. Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023.

[15] J. Chung, S. Lee, H. Nam, J. Lee, and K. M. Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

[16] X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhende, X. Wang, A. Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[17] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[18] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[19] J. Dong and Y.-X. Wang. Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Q. Dong, C. Cao, and Y. Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022.

[21] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999.

[22] D. Epstein, A. Jabri, B. Poole, A. Efros, and A. Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36: 16222–16239, 2023.

[23] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

[26] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[27] A. Hertz, K. Aberman, and D. Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2328–2337, 2023.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[29] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[30] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[31] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

[32] B. Hu, J. Huang, Y. Liu, Y.-W. Tai, and C.-K. Tang. Nerf-rpn: A general framework for object detection in nerfs. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.02253. URL `http://dx.doi.org/10.1109/CVPR52729.2023.02253`.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[34] N. Kalischek, T. Peters, J. D. Wegner, and K. Schindler. Tetrahedral diffusion models for 3d shape generation. *arXiv preprint arXiv:2211.13220*, 2022.

[35] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[36] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[37] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`.

[38] S. W. Kim, B. Brown, K. Yin, K. Kreis, K. Schwarz, D. Li, R. Rombach, A. Torralba, and S. Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8496–8506, 2023.

[39] Y. Liang, X. Yang, J. Lin, H. Li, X. Xu, and Y. Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.

[40] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

[41] Y. Lin, H. Han, C. Gong, Z. Xu, Y. Zhang, and X. Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023.

[42] H.-K. Liu, I. Shen, B.-Y. Chen, et al. Nerf-in: Free-form nerf inpainting with rgb-d priors. *arXiv preprint arXiv:2206.04901*, 2022.

[43] K. Liu, F. Zhan, Y. Chen, J. Zhang, Y. Yu, A. E. Saddik, S. Lu, and E. Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. 2023.

[44] Z. Liu, Y. Feng, M. J. Black, D. Nowrouzezahrai, L. Paull, and W. Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023.

[45] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023.

[46] Z. Lu, J. Jiang, J. Huang, G. Wu, and X. Liu. Glama: Joint spatial and frequency loss for general image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1301–1310, 2022.

[47] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[48] N. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995.

[49] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023.

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

[51] A. Mirzaei, T. Aumentado-Armstrong, M. A. Brubaker, J. Kelly, A. Levinshtein, K. G. Derpanis, and I. Gilitschenski. Reference-guided controllable inpainting of neural radiance fields. In *ICCV*, 2023.

[52] A. Mirzaei, T. Aumentado-Armstrong, M. A. Brubaker, J. Kelly, A. Levinshtein, K. G. Derpanis, and I. Gilitschenski. Watch your steps: Local image and scene editing by text instructions. *arXiv preprint arXiv:2308.08947*, 2023.

[53] A. Mirzaei, T. Aumentado-Armstrong, K. G. Derpanis, J. Kelly, M. A. Brubaker, I. Gilitschenski, and A. Levinshtein. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *CVPR*, 2023.

[54] A. Mirzaei, R. De Lutio, S. W. Kim, D. Acuna, J. Kelly, S. Fidler, I. Gilitschenski, and Z. Gojcic. Reffusion: Reference adapted diffusion models for 3d scene inpainting. *arXiv preprint arXiv:2404.10765*, 2024.

[55] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[56] C. Mou, X. Wang, J. Song, Y. Shan, and J. Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023.

[57] C. Mou, X. Wang, J. Song, Y. Shan, and J. Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. *arXiv preprint arXiv:2402.02583*, 2024.

[58] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[59] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[60] M. Niemeyer and A. Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

[61] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[62] J. Park, G. Kwon, and J. C. Ye. Ed-nerf: Efficient text-guided editing of 3d scene using latent space nerf. *arXiv preprint arXiv:2310.02712*, 2023.

[63] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

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

[65] K. Prabhu, J. Wu, L. Tsai, P. Hedman, D. B. Goldman, B. Poole, and M. Broxton. Inpaint3d: 3d scene content generation using 2d inpainting diffusion. *arXiv preprint arXiv:2312.03869*, 2023.

[66] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister. Langsplat: 3d language gaussian splatting. *arXiv preprint arXiv:2312.16084*, 2023.

[67] Y. Qu, S. Dai, X. Li, J. Lin, L. Cao, S. Zhang, and R. Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. *arXiv preprint arXiv:2405.17596*, 2024.

[68] A. Raj, S. Kaza, B. Poole, M. Niemeyer, N. Ruiz, B. Mildenhall, S. Zada, K. Aberman, M. Rubinstein, J. Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2349–2359, 2023.

[69] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[70] J. Ren, L. Pan, J. Tang, C. Zhang, A. Cao, G. Zeng, and Z. Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

[71] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[72] V. Rudnev, M. Elgharib, W. Smith, L. Liu, V. Golyanik, and C. Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision (ECCV)*, 2022.

[73] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[74] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

[75] A. Sargsyan, S. Navasardyan, X. Xu, and H. Shi. Mi-gan: A simple baseline for image inpainting on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7335–7345, 2023.

[76] E. Sella, G. Fiebelman, P. Hedman, and H. Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 430–440, 2023.

[77] P. Shamsolmoali, M. Zareapoor, and E. Granger. Transinpaint: Transformer-based image inpainting with context adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 849–858, 2023.

[78] Y. Siddiqui, A. Alliegro, A. Artemov, T. Tommasi, D. Sirigatti, V. Rosov, A. Dai, and M. Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. *arXiv preprint arXiv:2311.15475*, 2023.

[79] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[80] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[81] C. Sun, M. Sun, and H.-T. Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022.

[82] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[83] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023.

[84] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[85] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024.

[86] L. Tang, N. Ruiz, Q. Chu, Y. Li, A. Holynski, D. E. Jacobs, B. Hariharan, Y. Pritch, N. Wadhwa, K. Aberman, et al. Realfill: Reference-driven generation for authentic image completion. *arXiv preprint arXiv:2309.16668*, 2023.

[87] M. A. Uy, R. Martin-Brualla, L. Guibas, and K. Li. Scade: Nerfs from space carving with ambiguity-aware depth estimates. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[88] M. A. Uy, R. Martin-Brualla, L. Guibas, and K. Li. Scade: Nerfs from space carving with ambiguity-aware depth estimates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16518–16527, 2023.

[89] A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, K. Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022.

[90] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[91] D. Wang, T. Zhang, A. Abboud, and S. Süsstrunk. Inpaintnerf360: Text-guided 3d inpainting on unbounded neural radiance fields. *arXiv preprint arXiv:2305.15094*, 2023.

[92] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.

[93] E. Weber, A. Holynski, V. Jampani, S. Saxena, N. Snavely, A. Kar, and A. Kanazawa. Nerfiller: Completing scenes via generative 3d inpainting. In *CVPR*, 2024.

[94] S. Weder, G. Garcia-Hernando, Á. Monszpart, M. Pollefeys, G. Brostow, M. Firman, and S. Vicente. Removing objects from neural radiance fields. In *CVPR*, 2023.

[95] H. Weng, T. Yang, J. Wang, Y. Li, T. Zhang, C. Chen, and L. Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023.

[96] J. Wu, J.-W. Bian, X. Li, G. Wang, I. Reid, P. Torr, and V. A. Prisacariu. Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing. *arXiv preprint arXiv:2403.08733*, 2024.

[97] R. Wu, B. Mildenhall, P. Henzler, K. Park, R. Gao, D. Watson, P. P. Srinivasan, D. Verbin, J. T. Barron, B. Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. *arXiv preprint arXiv:2312.02981*, 2023.

[98] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[99] J. Ye, P. Wang, K. Li, Y. Shi, and H. Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023.

[100] M. Ye, M. Danelljan, F. Yu, and L. Ke. Gaussian grouping: Segment and edit anything in 3d scenes. *arXiv preprint arXiv:2312.00732*, 2023.

[101] T. Yi, J. Fang, J. Wang, G. Wu, L. Xie, X. Zhang, W. Liu, Q. Tian, and X. Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024.

[102] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

[103] J. Zhang, Z. Tang, Y. Pang, X. Cheng, P. Jin, Y. Wei, W. Yu, M. Ning, and L. Yuan. Repaint123: Fast and high-quality one image to 3d generation with progressive controllable 2d repainting. *arXiv preprint arXiv:2312.13271*, 2023.

[104] L. Zhang, Y. Zhou, C. Barnes, S. Amirghodsi, Z. Lin, E. Shechtman, and J. Shi. Perceptual artifacts localization for inpainting. In *European Conference on Computer Vision*, pages 146–164. Springer, 2022.

[105] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[106] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[107] M. Zheng, Z. Haiyu, H. Yang, and D. Huang. Neuface: Realistic 3d neural face rendering from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[108] Y. Zheng, X. Li, K. Nagano, S. Liu, O. Hilliges, and S. De Mello. A unified approach for text-and image-guided 4d scene generation. *arXiv preprint arXiv:2311.16854*, 2023.

[109] Q. W. K. L. C. J. X. K. Z. N. X. Y. L. Y. S. Y. C. Zhiheng Liu, Hao Ouyang. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024.

[110] J. Zhuang, C. Wang, L. Lin, L. Liu, and G. Li. Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023.

A Supplemental Material for "In-N-Out: Lifting 2D Diffusion Prior for 3D Object Removal via Tuning-Free Latents Alignment"

A.1 Implementation Detail

For the inpainting network, we employ the stable-diffusion-2-inpainting model [71], which encodes a masked image into the same dimensional latent space and integrates conditions via concatenation. We set the denoising steps for inpainting at 20. To achieve better generalization, we propose sampling the base frame according to the geometrical centroid of the training camera poses, meaning the camera that sits most centrally among the training views. However, we found that Stable Diffusion occasionally inpaints strange artifacts in the masked region. To mitigate this, we propose sampling n candidate views around the geometrical centroid and selecting the one with the highest similarity votes. This approach automatically avoids such occasion artifacts without human intervention. In our implementation, we used five candidate views, and the similarity was calculated using perceptual hashing. In the reprojection procedure of ELA, we adjust the camera intrinsics to match the latent dimensionality. Furthermore, to refine the ILA mechanism, we incorporate Cross-View Attention (CVA) into every self-attention layer of the inpainting model. Each step in this modified approach is controlled with λ_a set at 0.2.

For our 3D representation (NeRF) implementation, we utilize the "nerfacto" framework proposed by NerfStudio [83]. To ensure stable training, we deactivated the view-dependent effect. We pre-train the NeRF using 10000 iterations in stage 1 and jointly optimize it using 5000 iterations in stage 3. Our monocular depth estimation adopts DepthAnything [98], complemented by the depth loss outlined in DS-NeRF [17]. Moreover, we employ StyleGAN2 discriminator [35] to implement adversarial loss.

A.2 Sensitivity Analysis

We conducted several sensitivity analyses regarding the base view selection, λ_a in ILA, and the subset selection. Due to the computational burden, we conduct the sensitivity analysis on six out of ten scenes with higher inpainting variability from the SPIn-NeRF dataset.

(a) Base View Selection:

To achieve better generalization, we propose sampling the base frame according to the geometrical centroid of the training camera poses, meaning the camera that sits most centrally among the training views. However, we found that Stable Diffusion occasionally inpaints strange artifacts in the masked region. To mitigate this, we propose sampling n candidate views around the geometrical centroid and selecting the one with the highest similarity votes. This approach automatically avoids such occasion artifacts without human intervention. In our implementation, we used five candidate views, and the similarity was calculated using perceptual hashing.

We tested our results under different settings (candidate numbers): 3, 5, 7, and 9. The base frame selection algorithm proved to be robust, with our algorithm typically yielding the same base frame. However, another factor influencing this step is the random seed. Setting different seeds causes the 2D inpainting model to produce different results, leading to different base frames being selected. We tested our methods under five different seeds, and the final scores are reported in Table 4. While different seeds cause the final NeRF to differ in the appearance of the masked region, the consistency of the multi-view inpainting results remains robust, resulting in minimal variance in the evaluation scores.

(b) λ_a in ILA:

To effectively examine the effect of the hyper-parameter λ_a in ILA, we evaluated our method's rendering quality with different λ_a values of 0.2, 0.4, 0.6, and 0.8. The metrics are reported in Table 5. Quantitatively, the results are consistent across different λ_a values, indicating that the effect of this hyper-parameter is relatively small. This conclusion is also supported by qualitative results. Larger λ_a values tend to produce slight variations in some small regions, but the global structure and semantics are preserved. This stability is attributed to the significant role of the initial latent alignment in ELA, which effectively aligns the underlying inpainting structure, thereby maintaining low variability in appearance. Additionally, the self-attention layer, where cross-view attention is introduced, does not dominate the entire Stable Diffusion Unet. It is balanced by the presence of

Table 4: Sensitivity analysis on the prior inpainting results and prior view selection. Results are evaluated on the SPIn-NeRF dataset with different random seeds.

Seed	LPIPS ↓	MUSIQ ↑	FID ↓
1	0.46	46.61	264.91
2	0.44	48.04	255.29
3	0.44	46.47	262.09
4	0.44	45.72	261.04
5	0.46	48.65	258.50
Avg	0.45	47.10	260.37
Std	0.01	1.21	3.657

other (residual and linear) layers, ensuring cross-view attention does not override the signal during the denoising process. Hence we simply set λ_a as 0.2 in our implementation.

Table 5: Sensitivity analysis on λ_a used in ILA.

λ_a	LPIPS ↓	MUSIQ ↑	FID ↓
0.2	0.44	**47.11**	**261.62**
0.4	**0.44**	46.76	264.91
0.6	0.44	46.47	264.37
0.8	0.45	46.33	265.10
Avg	0.44	46.67	264.00
Std	0.01	0.35	1.62

(c) Subset Selection:

We found that for reconstruction tasks, more views can enhance quality; however, for generation tasks, using the entire set of images can introduce unnecessary inconsistencies. Therefore, we propose selecting the subset according to the distribution of camera viewpoints.

We evenly split the viewpoints into 12 groups based on the base view's camera space (evenly 2 on the x and y axes and 3 on the z-axis) and select 50 percent within each group according to perceptual hashing similarity to the base view. This approach avoids redundant views introducing supervision conflicts while covering different viewpoints for effective supervision.

We also evaluated our method based on different percentages, as reported in Table 6. The quantitative scores are quite close, indicating that for most scenes, the difference isn't significant. For one complex scene with extremely high frequencies, setting the percentage too low (0.2) yields artifacts in the test view due to insufficient viewpoint coverage. Conversely, setting the percentage too high (0.8) introduces appearance conflicts due to the high variability of the inpainted results.

Overall, for most scenes, the subset selection algorithm is robust due to the consideration of viewpoints distribution. For extreme cases, careful selection of the percentage might be necessary. However, values between 0.5 and 0.7 remain a reliable choice.

Table 6: Sensitivity analysis on proportion of images selected for the subset.

Percentage	LPIPS ↓	MUSIQ ↑	FID ↓
0.2	0.46	45.98	265.48
0.4	<u>0.44</u>	46.32	264.91
0.6	**0.44**	**47.11**	**261.62**
0.8	0.45	<u>46.47</u>	<u>263.20</u>

(d) λ_{patch} in patch loss:

To assess the sensitivity of the patch loss multiplier λ_{patch}, we evaluated the method's performance using various values of λ_{patch}: 0.001, 0.005, 0.01, 0.05, and 0.1. The results are reported in Table 7. Analysis of the table indicates that varying λ_{patch} leads to similar performance across different

settings, with a low standard deviation of the metrics. However, there is an observable trend where setting λ_{patch} too low or too high adversely affects performance. The multiplier λ_{patch} is critical as it determines the extent of influence multi-view images have on the NeRF. Insufficient multi-view supervision can lead to inadequate training, whereas excessive supervision may result in conflicting inputs. Consequently, we have set λ_{patch} at 0.01 in our implementation for optimal balance.

Table 7: Sensitivity analysis on λ_{patch} used for patch loss.

λ_{patch}	LPIPS ↓	MUSIQ ↑	FID ↓
0.001	0.46	46.078	263.32
0.005	0.45	47.08	262.43
0.010	**0.44**	**47.11**	**261.62**
0.050	0.47	44.93	265.31
0.100	0.49	44.05	277.36
Avg	0.46	45.85	266.01
Std	0.02	1.35	6.49

A.3 More Qualitative Results

This section presents extended qualitative results from our experiments on the SPIn-NeRF Dataset. Fig. 7 and Fig. 8 showcase a series of multi-view comparative inpaintings.

A.4 Details on User Study and Impact

To comprehensively evaluate our method using human subjects, we conducted a user study focusing on three aspects: (1) Background Coherence — assessing whether the inpainted area blends seamlessly with the remaining background, (2) Detail Preservation — determining if the inpainted area retains high-fidelity details, and (3) Overall Quality — gauging participants' preference rates for the inpainted results. For each method, we presented users with two multi-view test images from each scene and instructed them to choose the method that best met the criteria for each aspect. Clear instructions were provided to ensure participants understood the rating process. An example screenshot of the study interface is shown in Fig. 9.

The user study we conducted focused solely on collecting participants' preferences regarding different inpainting results, involving no sensitive or personal data collection beyond their aesthetic judgments. The study's design was inherently low-risk as it required participants to simply view and evaluate digital images based on their visual appeal and perceived quality. Furthermore, the participation was entirely voluntary, with clear instructions provided, allowing participants to withdraw at any time without any consequence. Given these factors, the potential for harm or discomfort to participants was negligible, ensuring the study maintained a minimal risk profile.

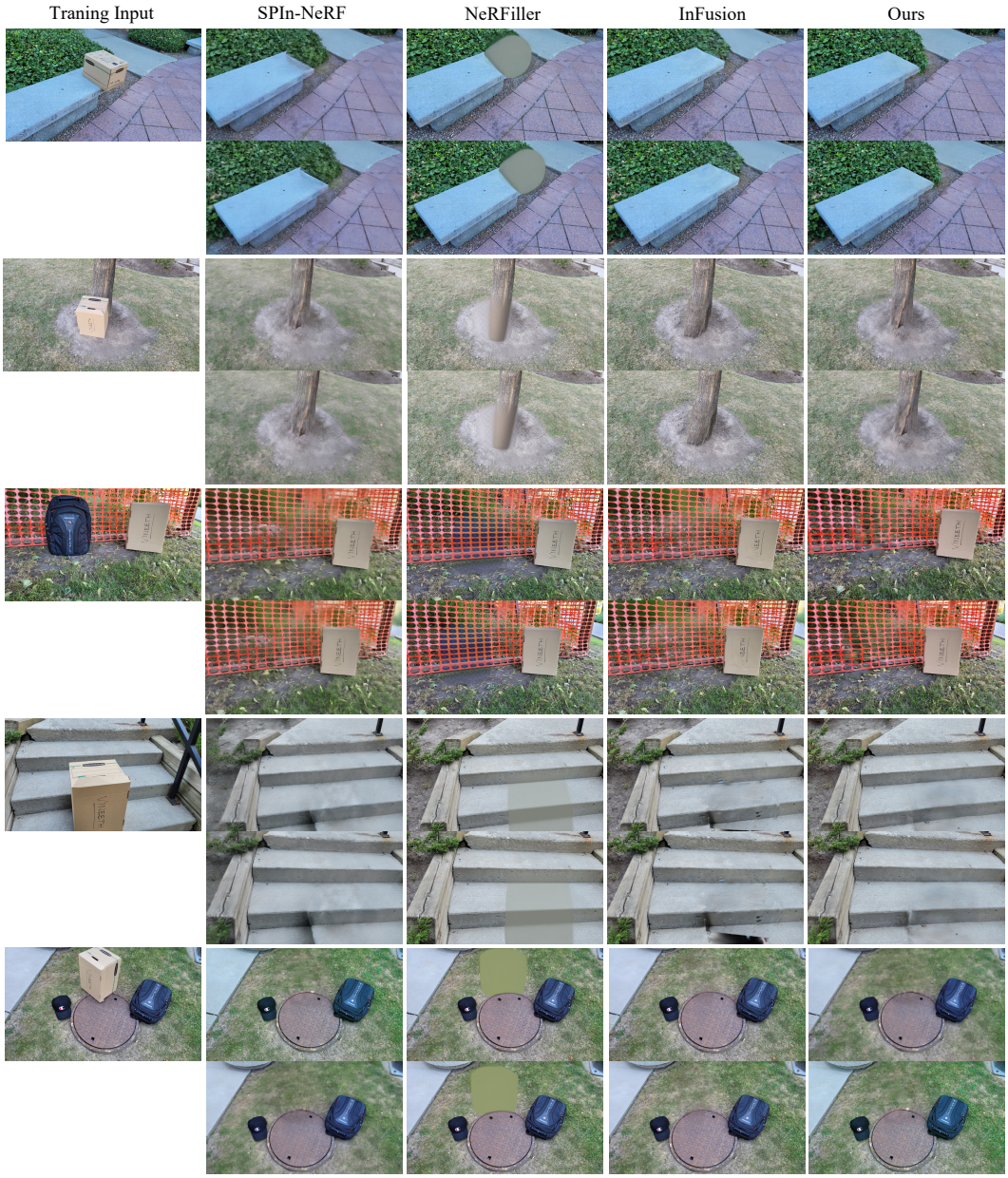


Figure 7: Additional Qualitative Results on the SPIn-NeRF Dataset.

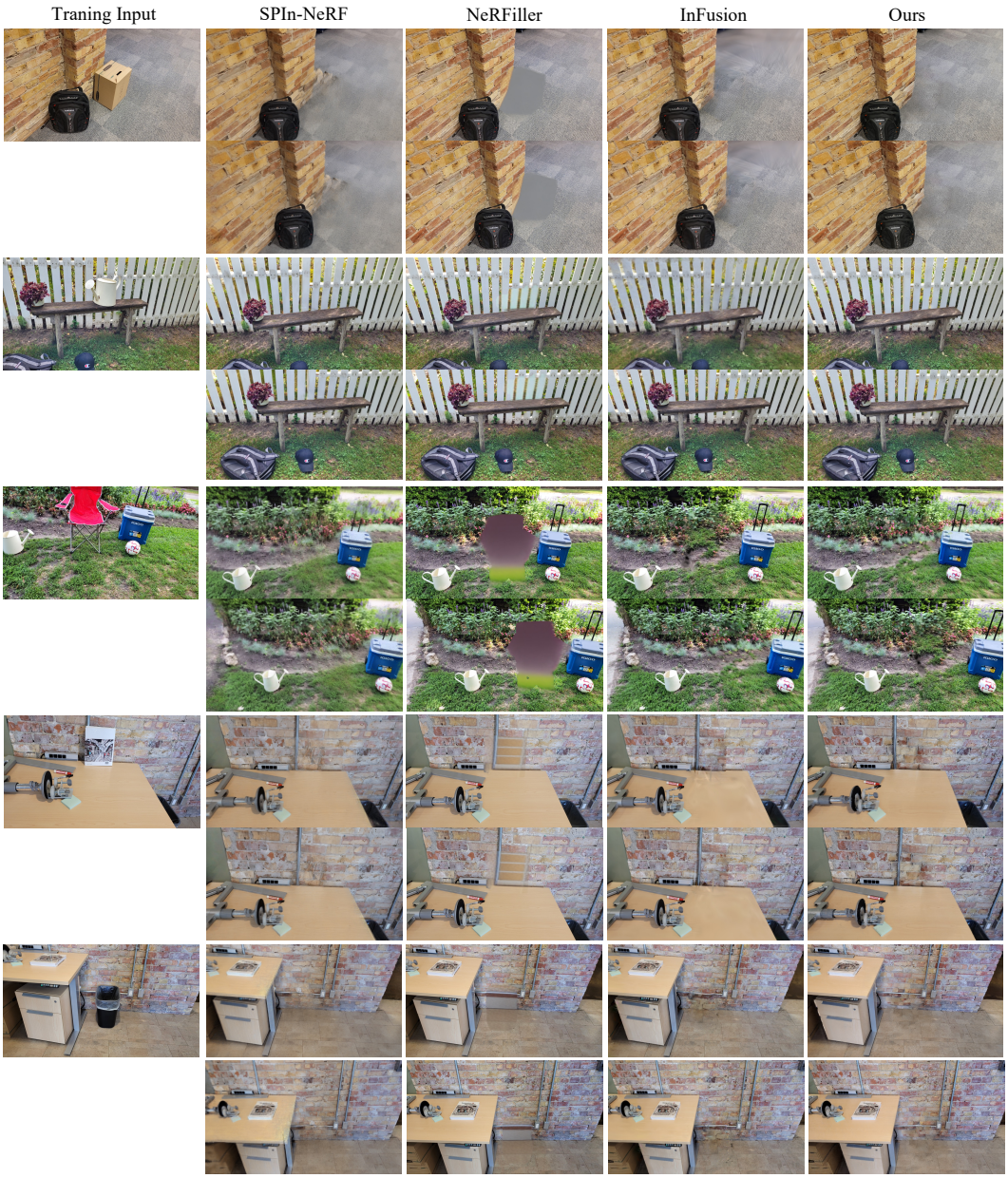


Figure 8: Multi-view Qualitative Results on the SPIn-NeRF Dataset.

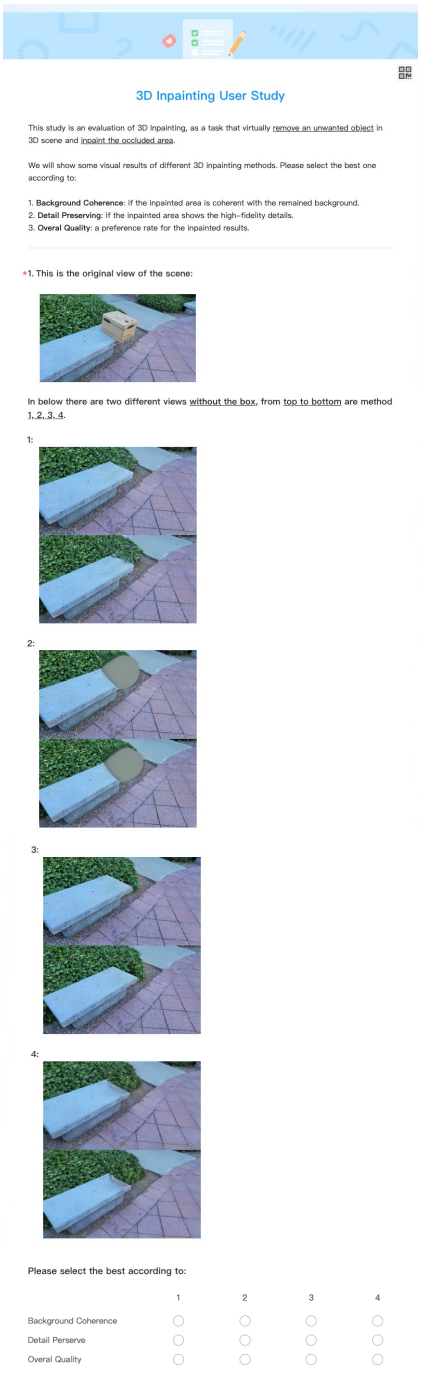


Figure 9: Example of User Study.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we clearly state that our method contributes to the field of object removal in neural radiance fields, offering a novel and effective solution to improve multi-view consistency.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the conclusion (Sec. 6) , we discuss the limitations of our approach, outlining the current bottlenecks.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper does not present theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We detail our implementation in Sec. A.1. The evaluation settings are clearly described in Sec. 5.1, ensuring a fair comparison by standardizing the input across all baselines.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Due to the extensive work required for this task, we will need to clean up the code before its release. However, to ensure reproducibility, we provide detailed instructions in Sec. 4 and Sec. A.1. The code will be made available upon completion of the cleanup process.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We clearly stated the evaluation setting and implementation details in Sec. 5.1 and Sec. A.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Previous 3D object removal methods did not report error bars, and the computational resources required for these experiments were relatively high, making it impractical to replicate the experiments multiple times.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided details on the computational hardware used in experiment (Sec. 5.1 and Sec. A.1).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and ensured that our research adheres to its guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the conclusion (Sec. 6), we discuss the positive societal impacts of our research and state the potential harms of technology misuse.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No new dataset or model is proposed in this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in this work are properly credited and are the license and terms of use explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We have provided clear instructions for human evaluation in our user study and included a sample screenshot in the supplementary materials, presented in Sec. 5 and Sec. A.4.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We have explicitly informed all participants in the user study that the collected data will be used solely for research purposes.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.