FreeFix: Boosting 3D Gaussian Splatting via Fine-Tuning-Free Diffusion Models

Hongyu Zhou[1], Zisen Shao[2], Sheng Miao[1], Pan Wang[3], Dongfeng Bai[3]
Bingbing Liu[3], Yiyi Liao[1]✉
[1] Zhejiang University [2] University of Maryland, College Park [3] Huawei

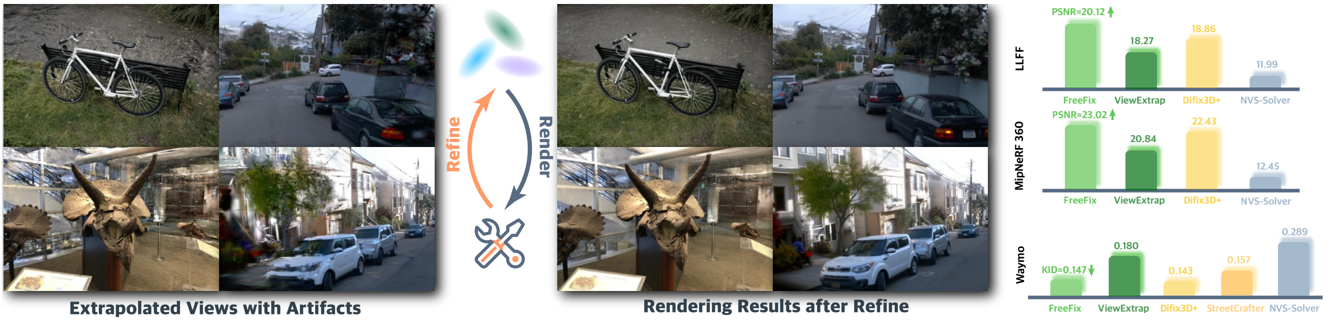


Figure 1. **Overview of FreeFix.** We present FreeFix, a method designed to improve the rendering results of extrapolated views in 3D Gaussian Splatting, without requiring fine-tuning of diffusion models. Experiments on multiple datasets show that FreeFix provides performance that is comparable to, or even superior to, most advanced methods that require fine-tuning.

Abstract

Neural Radiance Fields and 3D Gaussian Splatting have advanced novel view synthesis, yet still rely on dense inputs and often degrade at extrapolated views. Recent approaches leverage generative models, such as diffusion models, to provide additional supervision, but face a trade-off between generalization and fidelity: fine-tuning diffusion models for artifact removal improves fidelity but risks overfitting, while fine-tuning-free methods preserve generalization but often yield lower fidelity. We introduce FreeFix, a fine-tuning-free approach that pushes the boundary of this trade-off by enhancing extrapolated rendering with pretrained image diffusion models. We present an interleaved 2D–3D refinement strategy, showing that image diffusion models can be leveraged for consistent refinement without relying on costly video diffusion models. Furthermore, we take a closer look at the guidance signal for 2D refinement and propose a per-pixel confidence mask to identify uncertain regions for targeted improvement. Experiments across multiple datasets show that FreeFix improves multi-frame consistency and achieves performance comparable to or surpassing fine-tuning-based methods, while retaining strong generalization ability. Our project page is at https://xdimlab.github.io/freefix.

1. Introduction

Novel view synthesis (NVS) is a fundamental problem in 3D computer vision, playing an important role in advancing mixed reality and embodied artificial intelligence. Neural Radiance Fields (NeRF) [19] and 3D Gaussian Splatting (3DGS) [9] have achieved high-fidelity rendering, with 3DGS in particular becoming the mainstream choice for its real-time rendering capability. However, both methods require densely captured training images, which are often difficult to obtain, and they tend to produce artifacts at extrapolated viewpoints, namely those outside the interpolation range of the training views. These limitations hinder their use in downstream applications such as autonomous driving simulation and free-viewpoint user experiences.

Recent work has explored addressing artifacts in extrapolated view rendering with 3DGS. Existing approaches fall into two categories: adding regularization terms during training or augmenting supervision views using generative models. The regularization terms are often derived from 3D priors [10, 33, 48, 50, 52], or additional sensors [21], but they are typically hand-crafted and limited to specific scene types. Moreover, their lack of hallucination capability further restricts their applicability. In leveraging diffusion models (DMs), some approaches fine-tune them with paired data, e.g., by using sparse LiDAR inputs or extrapolated renderings with artifacts to generate refined images. Many of these methods train on domain-specific datasets, such as those for autonomous driving [20, 35, 36, 41], which inevitably compromises the generalization ability of DMs. More recently, Difix3D+ [37] fine-tunes SD Turbo [25] on a wider range of 3D datasets, improving generalization. However, the substantial effort required to curate 3D data and the high fine-tuning cost make this approach time-consuming and expensive to extend to other DMs. An alternative line of work seeks to improve extrapolated rendering without fine-tuning, typically by providing extrapolated renderings as guidance during the denoising step. This preserves the generalization capacity of DMs trained on large-scale data, but such methods still lag behind fine-tuned approaches that are specifically adapted to the task.

Given the generalization–fidelity trade-off, we ask: can extrapolated view rendering be improved with DMs without sacrificing generalization? To address this challenge, we focus on fine-tuning-free methods and enhance their effectiveness for NVS extrapolation. This is achieved with our proposed *2D–3D interleaved refinement strategy* combined with *per-pixel confidence guidance for fine-tuning-free image refinement*. Specifically, given a trained 3DGS, we sample an extrapolated viewpoint, render the 2D image, refine it with a 2D image diffusion model (IDMs), and integrate the refined image back into the 3D scene by updating the 3DGS before proceeding to the next viewpoint. This interleaved 2D-3D refinement ensures that previously enhanced views inform subsequent 2D refinements and improve multi-view consistency. Importantly, we introduce a confidence-guided 2D refinement, where a per-pixel confidence map rendered from the 3DGS highlights regions requiring further improvement by the 2D DM. This contrasts with previous training-free methods that rely solely on rendering opacity, leaving the DM to identify artifact regions on its own. While our confidence guidance could in principle be applied to video diffusion models (VDMs), advanced video backbones are typically more computationally expensive and use temporal down-sampling, which prevents the direct use of per-pixel guidance. We show that our 2D–3D interleaved optimization strategy achieves consistent refined images without relying on VDMs.

Our contribution can be summarized as follows: 1) We propose a simple yet effective approach for enhancing extrapolated 3DGS rendering without the need for fine-tuning DMs, featuring a 2D–3D interleaved refinement strategy and per-pixel confidence guidance. 2) Our method is compatible with various DMs and preserves generalization across diverse scene contents. 3) Experimental results demonstrate that our approach significantly outperforms existing fine-tuning-free methods and achieves comparable or even superior performance to training-based methods.

2. Related Work

Numerous works have made efforts on improving quality of NVS. In this section, we will discuss related works in NVS and 3D reconstruction. Furthermore, we will explore efforts that improve NVS quality by incorporating priors from geometry, physics or generative models.

Novel View Synthesis: NVS aims to generate photorealistic images of a scene from novel viewpoints. Early methods primarily relied on traditional image-based rendering techniques, such as Light Field Rendering [14], Image-Based Rendering [28], and Multi-Plane Image [30, 55]. These approaches typically interpolate between existing views and are often limited by dense input imagery and struggle with complex occlusions. The advent of deep learning revolutionized NVS, led by two major paradigms: NeRF [19] and 3DGS [9]. NeRF implicitly represents a scene and achieves high-quality results, but its training and rendering speeds are slow. In contrast, 3DGS offers rapid training and real-time rendering. However, a significant limitation of 3DGS is the occurrence of visual artifacts in extrapolated views, which are viewpoints far from the training data. These artifacts compromise the realism and geometric fidelity of the synthesized images. Mitigating these artifacts is the focus of this paper.

NVS with Geometry Priors: To enhance the robustness of NVS models and reduce reconstruction ambiguity, many works have introduced geometry priors. These priors provide key information about the scene's 3D structure, which can be explicitly provided by external sensors like LiDAR or depth cameras [8, 17, 21, 23, 36, 40, 41]. Other methods utilize strong structural priors often found in real-world scenes, such as the assumption that the ground is a flat plane [5, 10, 52], the sky can be modeled as a dome [4, 43], or that walls and tables in indoor scenes are predominantly orthogonal [48]. These structural assumptions help regularize the reconstruction process. While these geometry priors can mitigate some reconstruction challenges, they often fall short of completely solving the artifact problem in extrapolated views, especially when the initial geometric prior is itself inaccurate.

NVS with Generative Priors: Generative priors leverage pre-trained generative models to assist NVS tasks, particularly when dealing with data scarcity or missing information. Early works explored using Generative Adversarial Networks (GANs) to improve rendering quality [24, 26, 39], where the GAN's discriminator ensured the local realism of synthesized images. More recently, DMs [11–13, 22, 31, 32, 34, 42] have gained prominence for their powerful generative capabilities. Their application in NVS falls into two main categories. The first involves fine-tuning a pre-trained DM, which has learned powerful priors from datasets [35, 37, 38, 41, 47, 49, 54]. This process adapts the model's knowledge to scene-specific appearances but can be computationally expensive and time-consuming. The second category, which aligns with our proposed method, leverages a pre-trained DM as a zero-shot prior without fine-tuning. The key challenge here is determining what part of the rendered image should be used as guidance for

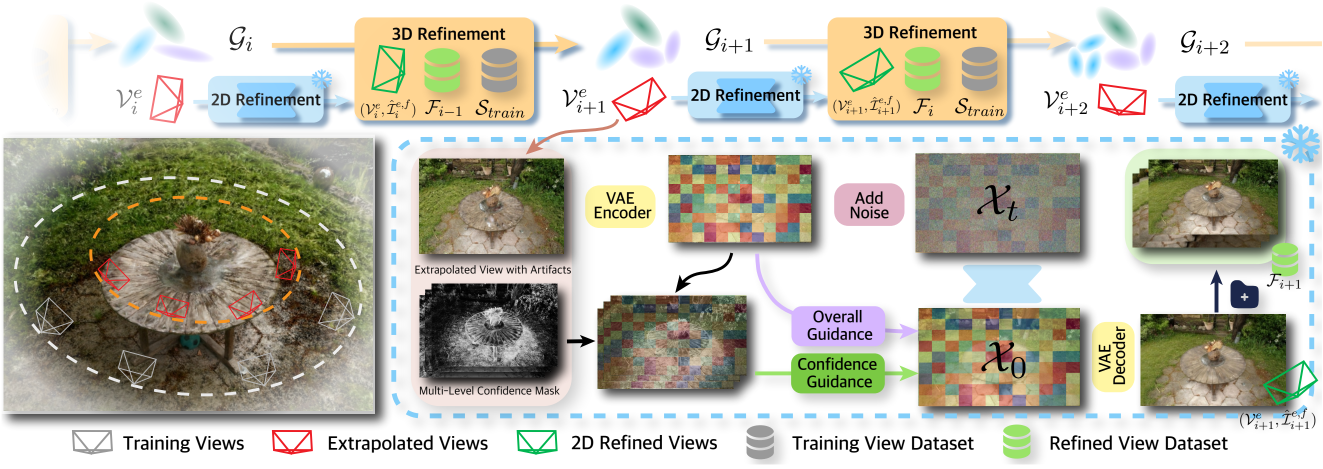


Figure 2. **Method.** FreeFix improves the rendering quality of extrapolated views in 3DGS without fine-tuning DMs, as illustrated in the bottom left of the pipeline. We propose an interleaved strategy that combines 2D and 3D refinement to utilize image diffusion models for generating multi-frame consistent results, as shown at the top of the pipeline. In the 2D refinement stage, we also introduce confidence guidance and overall guidance to enhance the quality and consistency of the denoising results.

the DM, and how to maintain multi-view consistency. Using the opacity channel of the rendered image as guidance is a common but often crude solution [16, 45, 46], as areas with high opacity can still be artifacts. Additionally, ensuring consistency across different novel views using IDMs is a critical problem. While VDMs [11, 31, 32, 42] can inherently handle this, they are often computationally heavy and not suitable for all applications.

3. Method

The FreeFix pipeline is illustrated in Fig. 2. In this section, we will first define our task and the relevant notations in Sec. 3.1. Next, we will introduce the interleaved refinement strategy for 2D and 3D refinement in Sec. 3.3. Finally, we will discuss the guidance utilized in diffusion denoising in Sec. 3.4.

3.1. Preliminaries

Task Definition: In the paper, we focus on the task of refining existing 3DGS. Specifically, given a 3DGS model $\mathcal{G}_{init}$ reconstructed from sparse view or partial observations $\mathcal{S}_{train} = \{(\mathcal{V}_0^t, \mathcal{I}_0^t), (\mathcal{V}_1^t, \mathcal{I}_1^t), ..., (\mathcal{V}_n^t, \mathcal{I}_n^t)\}$, artifacts tend to appear on the rendering results $\pi(\mathcal{V}_i^e; \mathcal{G}_{init})$, which are rendered from a continuous trajectory consisting of m extrapolated views $\mathcal{T}_{ext} = \{\mathcal{V}_0^e, \mathcal{V}_1^e, ..., \mathcal{V}_m^e\}$. Our objective is to fix these artifacts in the extrapolated views and refine the initial 3DGS into $\mathcal{G}_{refined}$. The extrapolated view rendering results from the refined 3DGS, $\pi(\mathcal{V}_i^e; \mathcal{G}_{refined})$, are expected to show improvements over the initial 3DGS results.

3D Gaussian Splatting: 3D Gaussian Splatting defines 3D Gaussians as volumetric particles, which are parameterized by their positions μ, rotations $\mathbf{q}$, scales $\mathbf{s}$, opacities η, and color $\mathbf{c}$. The covariance $\mathbf{\Sigma}$ of 3D Gaussians is defined as $\mathbf{\Sigma} = \mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{R}^T$, where $\mathbf{R} \in \mathbf{SO}(3)$ and $\mathbf{S} \in \mathbb{R}^{3\times3}$ represent the matrix formats of $\mathbf{q}$ and $\mathbf{s}$. Novel views can be rendered from 3DGS as follows:

$$\alpha_i = \eta_i \exp[-\frac{1}{2}(\mathbf{p}-\mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{p}-\mu_i)]$$
$$\pi(\mathcal{V};\mathcal{G}) = \sum_{i=1}^{N} \alpha_i \mathbf{c}_i \prod_{j}^{i-1}(1-\alpha_i) \tag{1}$$

Note that $\mathbf{c}_i$ can be replaced as other attributions to render additional modalities. For example, $\pi(\mathcal{V}; (\mathcal{G}, \mathbf{d}_i)) = \sum_{i=1}^{N} \alpha_i \mathbf{d}_i \prod_{j}^{i-1}(1-\alpha_i)$ denotes the rendering of a depth map, where $\mathbf{d}_i$ represents the depth of each Gaussian relative to viewpoint $\mathcal{V}$.

Diffusion Models: DMs generate a prediction $\hat{x}_0 \sim p_{data}$ that aligns with real-world distribution through iterative denoising. Specifically, the input of DMs is pure noise $\epsilon \sim \mathcal{N}(0, I)$ or real world data with added noise $x_t = (1-\sigma)x_0 + \sigma\epsilon$. DMs utilize a learnable denoising model $\mathbb{F}_\theta$ to minimize the denoising score matching objective:

$$\hat{x}_0^t = x_t - \sigma_t \mathbb{F}_\theta(x_t, t)$$
$$\mathbb{E}_{x_0,\epsilon,t}[||x_0 - \hat{x}_0^t||_2^2] \tag{2}$$

The next step denoising input x_{t-1} is derived as follows:

$$x_{t-1} = x_t + (\sigma_{t-1} - \sigma_t)\mathbb{F}_\theta(x_t, t) \tag{3}$$

The denoising step iterates until the prediction $\hat{x}_0$ is obtained.

3.2. Method Overview

DMs are powerful tools for improving 3D reconstruction results due to their ability to hallucinate contents. VDMs

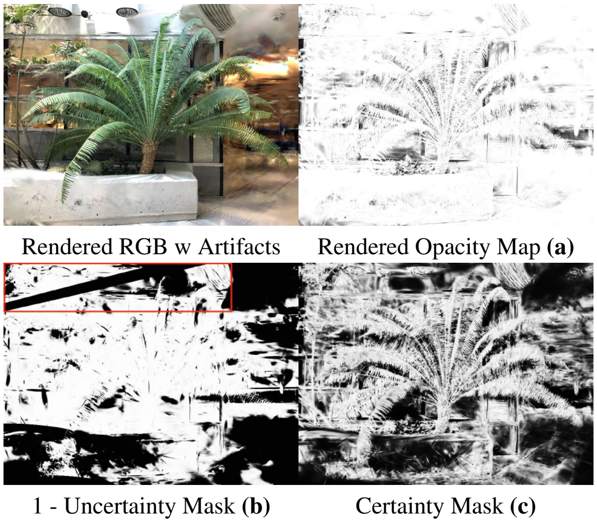


Figure 3. **Masks Comparison.** We aim to generate masks for guidance during denoising to fix artifacts in rendered RGBs. **(a)** Rendered opacity maps do not account for the presence of artifacts. **(b)** Uncertainty Masks are aware of artifacts; however, due to their numerical instability, the volume rendering processing can be *overwhelmed* by low-opacity Gaussians with large uncertainties. **(c)** The certainty mask we propose is numerically stable and robust against various types of artifacts.

are widely used for improving 3DGS [9] because of the inherent capability to apply attention across frames, ensuring multi-frame consistency. However, the temporal attention mechanism also introduces a computational burden, which also limits the output length of VDMs, as the computation complexity is quadratic in relation to the sequence length. Furthermore, recent advanced VDMs [11, 31, 42] utilize 3D VAE as their encoder and decoder, which performs temporal down-sampling, making it challenging to apply per-pixel confidence guidance.

Due to the above reasons, we select IDMs as the backbone in FreeFix. However, most existing IDMs are not designed for the novel view synthesis task and do not take reference views as input. IP-Adapter [44] accepts image prompts as input, but it is intended for style prompts rather than novel view synthesis. Directly applying IDMs can lead to inconsistency across frames and finally result in blurriness in refined 3DGS. To tackle the problem, we propose an interleaved refining strategy, multi-level confidence guidance, and overall guidance.

3.3. Interleaved Refinement Strategy

2D Refinement: As mentioned in Sec. 3.1, the trajectory of extrapolated views $\mathcal{T}_{ext} = \{\mathcal{V}_0^e, \mathcal{V}_1^e, ..., \mathcal{V}_m^e\}$ in our task definition is intended to be continuous. This continuous trajectory setting ensures that adjacent views $\mathcal{V}_i^e$ and $\mathcal{V}_{i+1}^e$ undergo only small transformations. A naive approach to keep consistency would be warping pixels from $\mathcal{V}_i^e$ to $\mathcal{V}_{i+1}^e$ and using DMs for inpainting. However, both rendered depth and predicted depth are not reliable for warping. Instead, we propose an interleaved refining strategy to enhance multi-view consistency.

Specifically, the refining process is interleaved and incremental along the trajectory $\mathcal{T}$. Given the current view $\mathcal{V}_i^e$, the current 3DGS $\mathcal{G}_{i-1}$ and rendered image $\hat{\mathcal{I}}_i^e = \pi(\mathcal{V}_i^e; \mathcal{G}_{i-1})$, we utilize denoising with guidance, as discussed in Sec. 3.4, to obtain the fixed image $\hat{\mathcal{I}}_i^{e,f}$. We also maintain a fixed image set $\mathcal{F}_{i-1} = \{(\mathcal{V}_0^e, \hat{\mathcal{I}}_0^{e,f}), (\mathcal{V}_1^e, \hat{\mathcal{I}}_1^{e,f}), ..., (\mathcal{V}_{i-1}^e, \hat{\mathcal{I}}_{i-1}^{e,f})\}$. We refine the current 3DGS $\mathcal{G}_{i-1}$ to $\mathcal{G}_i$ by using the training set $\mathcal{S}_{train}$, the previous refined view set $\mathcal{F}_{i-1}$ and the current refined image $\hat{\mathcal{I}}_i^{e,f}$.

3D Refinement: The supervision during 3D Refinement for $\mathcal{G}_i$ comes from current refined view $(\mathcal{V}_i^e, \hat{\mathcal{I}}_i^{e,f})$, $\mathcal{F}_{i-1}$ and S_{train}. The detailed sampling strategy for training is illustrated in the supplements.

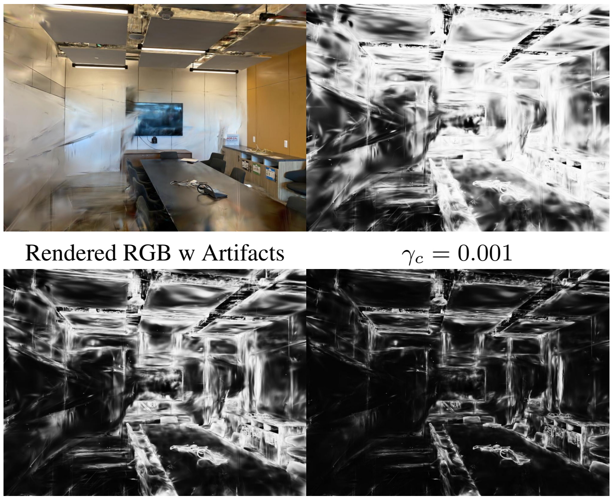
The generated results do not guarantee 3D consistency with training views, so we employ a smaller training loss for the generated views to prevent inaccurately generated areas from distorting 3D scenes. Additionally, the generated results exhibit slightly color bias compared to training views, which are often difficult for humans to distinguish. However, when applying the interleaved refining strategy, these slight color biases will accumulate, which may lead to a blurry and over-gray effect. We implement a simple yet efficient technique similar to [53] to tackle the problem. For each generated view, we define two optimizable affine matrices $\mathcal{A}_f \in \mathbb{R}^{3\times3}$ and $\mathcal{A}_b \in \mathbb{R}^{3\times1}$. The rendering results used for computing the training loss are applied to these affine matrices to avoid learning color bias:

$$\begin{aligned}
&\hat{\mathcal{I}}^{e'} = \mathcal{A}_f \times \hat{\mathcal{I}}^e + \mathcal{A}_b \\
&\mathcal{L} = (1-\lambda_s)||\hat{\mathcal{I}}^{e'} - \hat{\mathcal{I}}^{e,f}||_1 + \lambda_s SSIM(\hat{\mathcal{I}}^{'}, \hat{\mathcal{I}}^{e,f}) \qquad (4)
\end{aligned}$$

3.4. Denoising with Guidance

Given the rendered results of an extrapolated view, even though the image contains artifacts, most areas can still be regarded as photo-realistic rendering results. These regions with relatively high fidelity can provide essential information for generating an image free of artifacts, while maintaining almost the same content.

Experiments in Difix3D+ [37] have demonstrated that adding noise to images with artifacts and directly applying denoising using DMs can effectively remove these artifacts; however, the strength of the added noise is quite sensitive. For regions with significant artifacts, a larger scale of noise is needed to repaint those areas, while a smaller scale of noise is sufficient for areas with minimal artifacts. Although it may seem intuitive to apply different levels of noise to different regions, this approach does not align the



Rendered RGB w Artifacts

$\gamma_c = 0.001$

$\gamma_c = 0.01$

$\gamma_c = 0.1$

Figure 4. **Multi-Level Certainty Masks.** FreeFix employs multiple γ_c to obtain multi-level certainty masks as guidance. Each level of mask guides a different stage of denoising. A small γ_c with high overall certainty is used for the early stages of denoising, while a large γ_c which offers greater accuracy, is applied during the later stages of denoising.

data distribution of DMs. Instead, employing guidance during the diffusion denoising step is more practical and has been widely adopted in [16, 45].

Confidence Map: Utilizing appropriate guidance is an effective method for generating high-fidelity images while preserving accurate rendering results. However, current approaches that use warp masks or rendering opacities as guidance weights do not account for the presence of artifacts. For example, as illustrated in Fig. 3 (a), even when severe artifacts are present, the rendering opacities remain high, indicating that these artifacts continue to act as strong guidance during the denoising process. To tackle this issue, we propose utilizing confidence masks as guidance weights, as shown in Fig. 3 (c). The confidence scores are derived from Fisher information, which is also referenced in [6, 7]. Specifically, Fisher information measures the amount of information that the observation (x, y) carries about the unknown parameters w that model $p_f(y|x; w)$. In the context of novel view synthesis, Fisher information can be defined as:

$$p_f(\pi(\mathcal{V}; \mathcal{G})|\mathcal{V}; \mathcal{G}) \tag{5}$$

where $\mathcal{V}$ and $\mathcal{G}$ represent viewpoint and 3DGS respectively, while $\pi(\mathcal{V}; \mathcal{G})$ denotes the volume rendering results at the specific view $\mathcal{V}$.

The negative log likelihood of Fisher information in eq. (5), which serves as the uncertainty $\bar{\mathcal{C}}_{\mathcal{V};\mathcal{G}}$ of $\mathcal{G}$ at view $\mathcal{V}$, can be approximately derived as a Hessian matrix, the detailed derivation can be found in the supplementary materials:

$$\begin{aligned}\bar{\mathcal{C}}_{\mathcal{V};\mathcal{G}} &= -\log p_f(\pi|\mathcal{V}; \mathcal{G}) \\ &= \mathbf{H}^{''}[\pi|\mathcal{V}; \mathcal{G}] \\ &= \nabla_{\mathcal{G}}\pi(\mathcal{V}; \mathcal{G})^T \nabla_{\mathcal{G}}\pi(\mathcal{V}; \mathcal{G})\end{aligned} \tag{6}$$

[6, 7] renders the attribute $\bar{\mathcal{C}}_{\mathcal{V};\mathcal{G}}$ in volume rendering to obtain the uncertainty map. However, uncertainty is not a numerically stable representation, as its value can range from $[0, +\infty)$. As illustrated in Fig. 3 (b), the numeric instability of uncertainty may render an inaccurate uncertainty map. This often occurs when there are Gaussians with low opacity and high uncertainty, which can *overwhelm* the volume rendering. Instead, we use the complementary value as guidance, certainty $\mathcal{C}_{\mathcal{V};\mathcal{G}}$, also referred to as confidence in this paper, which has a stable numeric range of $[0, 1]$. The certainty $\mathcal{C}^{\gamma_c}_{\mathcal{V};\mathcal{G}}$ is defined as:

$$\mathcal{C}^{\gamma_c}_{\mathcal{V};\mathcal{G}} = \exp[-\gamma_c \bar{\mathcal{C}}_{\mathcal{V};\mathcal{G}}] \tag{7}$$

where γ_c is a hyperparameter. When $\gamma_c = 1$, we actually use the original Fisher information as the confidence. When render $\mathcal{C}_{\mathcal{V};\mathcal{G}}$ with hyperparameter as an attribute in 3DGS, and multiply with rendered opacity $\mathcal{M}^\alpha$, we obtain the confidence map $\mathcal{M}^{\gamma_c}_{\mathcal{V};\mathcal{G}}$:

$$\begin{aligned}\mathcal{M}^\alpha &= \pi(\mathcal{V}; (\mathcal{G}, \alpha)) \\ \mathcal{M}^{\gamma_c}_{\mathcal{V};\mathcal{G}} &= \pi(\mathcal{V}; (\mathcal{G}, \mathcal{C}^{\gamma_c}_{\mathcal{V};\mathcal{G}})) \odot \mathcal{M}^\alpha\end{aligned} \tag{8}$$

Multi-Level Confidence Maps: As shown in Fig. 4, γ_c is a hyperparameter that controls sensitivity to artifacts when rendering confidence maps. The larger the value of γ_c, the more sensitive the rendered confidence map becomes to artifacts. Selecting a single appropriate γ_c is not trivial. Therefore, we apply multi-level confidence maps as guidance. Since DMs generate a coarse structure of image rather than detailed appearance in the early denoising stages [27], we provide $\mathcal{M}^{\gamma_c}_{\mathcal{V};\mathcal{G}}$ with a small γ_c to offer more comprehensive guidance. In the later denoising stages, DMs tend to generate detailed appearances, so we provide $\mathcal{M}^{\gamma_c}_{\mathcal{V};\mathcal{G}}$ with a large γ_c to ensure that the guidance is sufficiently accurate.

Confidence Guidance: Given the rendered image $\hat{I}_{\mathcal{V};\mathcal{G}}$ and the corresponding confidence map $\mathcal{M}^{\gamma_c}_{\mathcal{V};\mathcal{G}}$, we can provide denoising guidance to DMs. We denote the rendered image after VAE encoding as x_0^r, and the resized confidence map that aligns with the shape of the latent space as $\mathcal{M}^c$. As illustrated in eq. (2), the predicted x_0^t at t timestep is given by $x_t - \sigma_t \mathbb{F}_\theta(x_t, t)$. We guide the model prediction as $x_0^{t,g}$ by blending the rendered image using confidence mask:

$$x_0^{t,g} = \mathcal{M}^c \odot x_0^r + (1 - \mathcal{M}^c) \odot x_0^t \tag{9}$$

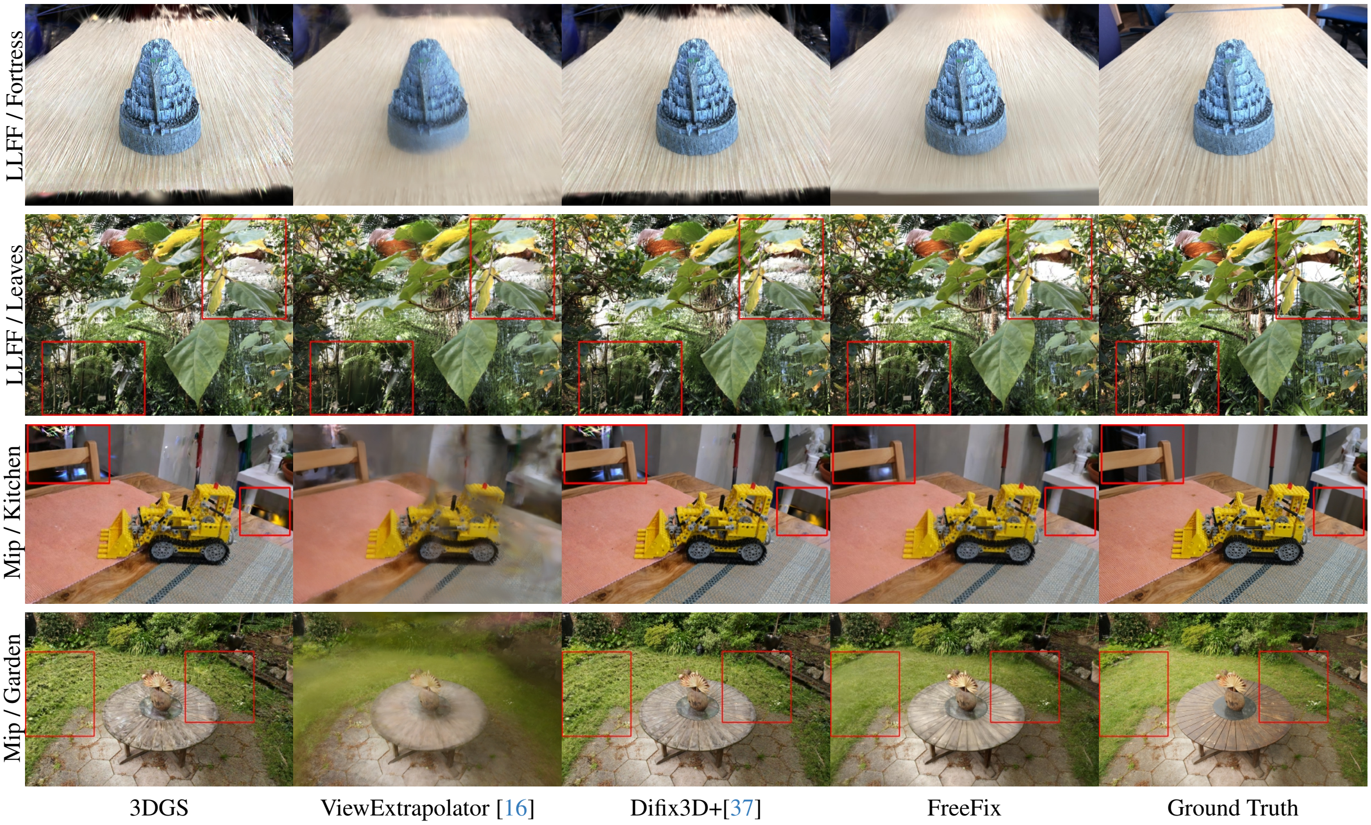


Figure 5. **Qualitative Comparisons on LLFF [18] and Mip-NeRF 360 [1].** FreeFix demonstrates state-of-the-art performance on these two datasets.

However, the input for the next denoising step cannot be directly obtained using eq. (3) since the model prediction x_0^t has been changed. Instead, we derive the new x_{t-1} by solving the following equations:

$$\begin{aligned} x_{t-1} &= x_0 + \sigma_{t-1}\mathbb{F}_\theta(x_t, t) \\ x_{t-1} &= x_t + (\sigma_{t-1} - \sigma_t)\mathbb{F}_\theta(x_t, t) \end{aligned} \tag{10}$$

The representation of x_{t-1} derived from $x_0^{t,g}$ and x_t is:

$$x_{t-1} = \frac{\sigma_{t-1}}{\sigma_t}x_t - \frac{\sigma_{t-1} - \sigma_t}{\sigma_t}x_0^{t,g} \tag{11}$$

Overall Guidance: Although the interleaved refining strategy provides higher fidelity rendering results and ensures that the rendering is more consistent with the generated content, using IDMs may still encounter issues of inconsistency in areas with low confidence. Particularly in regions with weak textures like ground and sky, the confidence map tends to be low, and allowing denoising to proceed freely in these areas can result in high inconsistency and blurriness in 3DGS. To address this issue, we propose an overall guidance approach, which combines confidence guidance in the very early stages of denoising to provide structural hints for the images. The combination of certainty and overall guidance is defined as follows:

$$\begin{aligned} x_0^{t,g} =& \mathcal{M}^c \odot x_0^r + \\ & (1 - \mathcal{M}^c) \odot (\beta\mathcal{M}^\alpha x_0^r + (1 - \beta\mathcal{M}^\alpha)x_0^t) \end{aligned} \tag{12}$$

where β is a hyperparameter that controls the strength of the overall guidance.

4. Experiments

Datasets: We conduct a series of experiments to evaluate the performance of FreeFix across multiple datasets with varying settings. We select LLFF [18] as the evaluation dataset for forward-facing scenes, Mip-NeRF 360 [1] for object-centric scenes, and Waymo [29] for driving scenes. For the LLFF and MipNeRF datasets, which contain relatively dense captured images, we select sparse or partially observed views as the training set and choose an extrapolated view trajectory that is distant from the views in the training set. The Waymo dataset only provides captured images from a single pass down the street, making it relatively sparse. We only utilize the front cameras as the training set and then translate or rotate the training cameras to create the test views. Details on the design of the training and testing views are provided in the supplementary materials.

	LLFF [18]			Mip-NeRF 360 [1]			Waymo [29]	DM Type	w/o Finetune	Only RGBs	3D Render
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	KID↓				
3DGS [9]	18.10	0.633	0.265	21.83	0.643	0.239	0.155	N/A	N/A	✓	✓
FreeFix + SDXL	19.93	0.695	0.237	22.68	0.685	0.213	0.150	Image	✓	✓	✓
FreeFix + Flux	20.12	0.700	0.221	23.02	0.689	0.208	0.147	Image	✓	✓	✓
ViewExtrapolator [16]	18.27	0.614	0.338	20.84	0.591	0.332	0.180	Video	✓	✓	✓
NVS-Solver [45]	11.99	0.351	0.560	12.45	0.266	0.631	0.289	Video	✓	✓	✗
Difix3D+ [37]	18.86	0.658	0.239	22.43	0.661	0.210	0.143	Image	✗	✓	✓
StreetCrafter [41]	N/A	N/A	N/A	N/A	N/A	N/A	0.157	Video	✗	✗	✓

Table 1. **Quantitative Comparison with Baselines.** FreeFix demonstrates superior performance among baselines without fine-tuning. Compared to models that require fine-tuning, FreeFix providing better results on LLFF and Mip-NeRF 360, while achieving comparable performance on Waymo. First , second , and third performances in each column are indicated by their respective colors.

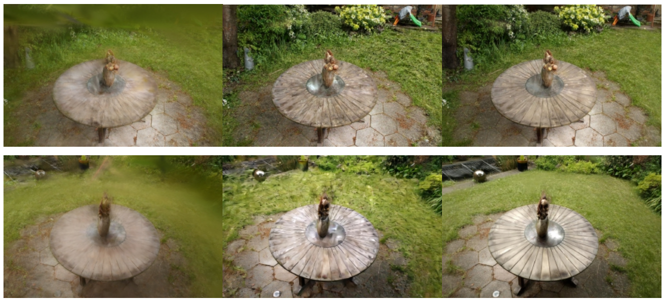

VE [16] + SVD FreeFix + SVD FreeFix + Flux

Figure 6. **Qualitative Ablation on Diffusion Models Selection.** FreeFix + Flux yields results with higher fidelity than FreeFix + SVD. Additionally, the improved results of FreeFix + SVD compared to ViewExtrapolator + SVD highlight the effectiveness of confidence guidance.

Model Settings and Baselines: FreeFix utilizes two powerful IDMs as its backbone: SDXL [22] and Flux [12], to showcase the capabilities of our method. For baseline selection, we consider various methods with different settings. For fine-tuning-free methods, we select ViewExtrapolator [16], and NVS-Solver [45] as the baseline. While ViewExtrapolator refines 3DGS with generated views like ours, NVS-Solver employs VDMs as the final renderer, without using 3D renderers, which consumes more computational resources during rendering. For methods that require fine-tuning of DMs, we choose Difix3D+ [37] and StreetCrafter [41] as baselines. StreetCrafter focuses on urban scenes and requires both LiDAR and RGB observations as input, while Difix3D+ is more generalizable and only requires RGB images. For all methods with a 3D renderer, we apply nearly the same 3D refining steps, ensuring that there are sufficient refining steps for the models to converge.

Evaluation Metrics: For the experiments on LLFF and MipNeRF, we adopt the most common settings for quantitative assessments, which include the evaluation of PSNR, SSIM, and LPIPS [51]. In the case of the Waymo dataset, where no ground truth is available for the test images, we utilize KID [2] for quantitative assessments.

4.1. Comparison with Baselines

We evaluate FreeFix using SDXL [22] and Flux [12] as the diffusion backbone on the LLFF, Mip-NeRF 360, and Waymo datasets. This includes a quantitative comparison in Tab. 1 and qualitative comparisons in Fig. 5 and Fig. 7 against baseline methods. Although FreeFix utilizes only IDMs as the backbone and does not require fine-tuning of the DMs, it still demonstrates performance that is comparable to, or even surpasses, methods that use VDMs or require fine-tuning, both in quantitative and qualitative assessments.

Specifically, ViewExtrapolator [16], which uses opacity masks as guidance, shows slight improvements in LLFF, although the improvement is less significant compared to our confidence-guided solution. Moreover, it fails to provide improvements in Mip-NeRF 360 and Waymo. This is due to the fact that ViewExtrapolator uses the nearest view from a set of training views as the reference view to generate the test views in a video diffusion model. While using the nearest training view as the reference view in SVD performs well in the forward-facing scenes in LLFF, where the test views are closer to the training views, this is usually not the case for Mip-NeRF 360 and Waymo, hence ViewExtrapolator yields degraded performance.

Difix3D+ demonstrates the most generalizability and powerful performance across our baselines. FreeFix surpasses Difix3D+ [37] in LLFF and Mip-NeRF 360, while providing comparable performance in Waymo. We attribute this to the generalizability of DMs. Although Difix3D+ is finetuned on DLV3D [15] and may have encountered similar scenes to those in LLFF and Mip-NeRF 360, the domain gap between datasets still weakens the generalizability of Difix3D+. In contrast, our method maintains the original generalizability of DMs learned from web-scale datasets. Regarding the Waymo dataset, Difix3D+ is fine-tuned on a large-scale in-house driving dataset, where driving scenes are highly structured and exhibit relatively small inter-class differences, making them easier for models to learn.

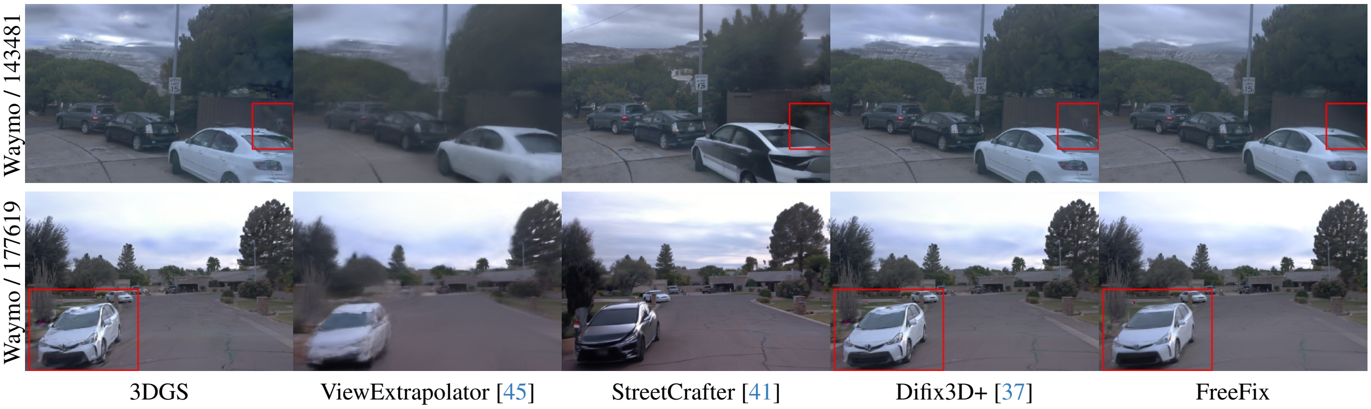


Figure 7. **Qualitative Comparisons on Waymo [29]**. FreeFix provide superior performance compared to ViewExtrapolator and StreetCrafter, and is comparable to Difix3D+ in the Waymo dataset. In some cases, FreeFix refines the scene even better than Difix3D+.

StreetCrafter [41] is tailored for urban scenes and requires LiDAR as input; for this reason, we only conduct experiments with this model on the Waymo dataset. In contrast to the original setting in StreetCrafter, our setup only provides the front camera to color the LiDAR points, which highlights the limitations of StreetCrafter in this context. NVS-Solver produces less satisfying results compared to other methods, which may be attributed to inaccurate depth estimation and warping results. We provide NVS-Solver results in supplementary materials.

Please note that we compute the average score across scenes for each dataset. We provide a quantitative comparison for each scene, along with additional qualitative comparisons in the supplementary materials.

4.2. Ablation Study

Image Diffusion Models vs Video Diffusion Models: FreeFix can also be applied to VDMs without temporal down-sampling, such as SVD [3]. Although SVD offers inherent consistency across frames, it suffers from blurriness compared to more advanced IDMs. We conduct an ablation study on the scene from MipNeRF-360/Garden to provide quantitative and qualitative comparisons in Tab. 2 and Fig. 6. Additionally, we include the results from ViewExtrapolator [16] on the same scene. While ViewExtrapolator also uses SVD as its backbone, it employs an opacity mask as guidance, which disentangles the effects of the differences in diffusion model backbones and helps demonstrate the effectiveness of our confidence guidance.

Effectiveness of Interleaved 2D-3D Refinement: The interleaved refining strategy, confidence guidance, and overall guidance are crucial for ensuring that the generation aligns with the original scenes and enhances consistency across frames. We conduct an ablation study of these modules on the scene from MipNeRF-360/Garden, as shown in Tab. 3. We perform experiments starting from a raw Flux model, which we slightly modify to function as an image-to-image model. We progressively add components from FreeFix to demonstrate the necessity of these techniques.

	PSNR↑	SSIM↑	LPIPS↓	Guidance
3DGS	18.38	0.415	0.357	N/A
VE [16] + SVD	17.86	0.409	0.505	Opacity
FreeFix + SVD	19.03	0.453	0.331	Certainty
FreeFix + SDXL	19.41	0.517	0.294	Certainty
FreeFix + Flux	**19.72**	**0.520**	**0.287**	Certainty

Table 2. **Quantitative Ablation on Diffusion Models Selection.**

	PSNR↑	SSIM↑	LPIPS↓
Raw Flux [12]	19.23	0.390	0.389
+ Confidence Guidance	19.32	0.435	0.349
+ Interleave Strategy	19.65	0.517	0.293
+ Overall Guidance	**19.72**	**0.520**	**0.287**

Table 3. **Ablation Study on Modules of FreeFix**. We incorporate each module from the raw Flux model to illustrate its necessity.

5. Conclusion

In this paper, we present FreeFix, a method for fixing artifacts and improving the quality of 3DGS without fine-tuning DMs. FreeFix demonstrates state-of-the-art performance across various datasets and possesses strong capabilities for deployment with future, more advanced DMs. However, FreeFix still has certain limitations. It may encounter failure cases when extrapolated views lead to excessive artifacts with minimal credible guidance. Additionally, the updating process for 3DGS is relatively slow and challenging to converge over dozens of refining steps. These challenges suggest opportunities for future work on designing more robust and efficient methods for integrating 3D reconstruction with 2D generative models.

Acknowledgements: This work is supported by NSFC under grant 62202418, U21B2004, and 62441223, the National Key R&D Program of China under Grant 2021ZD0114501, and Scientific Research Fund of Zhejiang University grant XY2025028.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 6, 7, 1

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 8

[4] Yun Chen, Jingkang Wang, Ze Yang, Sivabalan Manivasagam, and Raquel Urtasun. G3r: Gradient guided generalizable reconstruction. In *European Conference on Computer Vision*, pages 305–323. Springer, 2024. 2

[5] Zhiheng Feng, Wenhua Wu, and Hesheng Wang. Rogs: Large scale road surface reconstruction based on 2d gaussian splatting. *arXiv e-prints*, pages arXiv–2405, 2024. 2

[6] Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, and Tom Goldstein. Pup 3d-gs: Principled uncertainty pruning for 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5949–5958, 2025. 5

[7] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and mapping with radiance fields using fisher information. In *European Conference on Computer Vision*, pages 422–440. Springer, 2024. 5, 1

[8] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 2

[9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 4, 7

[10] Mustafa Khan, Hamidreza Fazlali, Dhruv Sharma, Tongtong Cao, Dongfeng Bai, Yuan Ren, and Bingbing Liu. Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction. *arXiv preprint arXiv:2407.02598*, 2024. 1, 2

[11] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 3, 4

[12] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 7, 8

[13] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2

[14] Marc Levoy and Pat Hanrahan. Light field rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 441–452. 2023. 2

[15] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 7

[16] Kunhao Liu, Ling Shao, and Shijian Lu. Novel view extrapolation with video diffusion priors. *arXiv preprint arXiv:2411.14208*, 2024. 3, 5, 6, 7, 8, 2, 4

[17] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 2

[18] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 6, 7, 1

[19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[20] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1559–1569, 2025. 1

[21] Yue Pan, Xingguang Zhong, Liren Jin, Louis Wiesmann, Marija Popović, Jens Behley, and Cyrill Stachniss. Pings: Gaussian splatting meets distance fields within a point-based implicit neural map. *arXiv preprint arXiv:2502.05752*, 2025. 1, 2

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 7

[23] Kevin Raj, Christopher Wewer, Raza Yunus, Eddy Ilg, and Jan Eric Lenssen. Spurfies: Sparse-view surface reconstruction using local geometry priors. In *International Conference on 3D Vision 2025*, 2025. 2

[24] Pierluigi Zama Ramirez, Diego Martin Arroyo, Alessio Tonioni, and Federico Tombari. Unsupervised novel view synthesis from a single image. *arXiv preprint arXiv:2102.03285*, 2021. 2

[25] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

[26] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in neural information processing systems*, 33:20154–20166, 2020. 2

[27] Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. Flowmo: Variance-based flow guidance for coherent motion in video generation. *arXiv preprint arXiv:2506.01144*, 2025. 5

[28] Heung-Yeung Shum, Shing-Chow Chan, and Sing Bing Kang. *Image-based rendering*. Springer, 2007. 2

[29] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 6, 7, 8, 1

[30] Richard Tucker and Noah Snavely. Single-View View Synthesis with Multiplane Images, 2020. arXiv:2004.11364. 2

[31] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 4

[32] Fusang Wang, Arnaud Louys, Nathan Piasco, Moussab Bennehar, Luis Roldão, and Dzmitry Tsishkou. PlaNeRF: SVD Unsupervised 3D Plane Regularization for NeRF Large-Scale Scene Reconstruction, 2023. arXiv:2305.16914 [cs]. 2, 3

[33] Fusang Wang, Arnaud Louys, Nathan Piasco, Moussab Bennehar, Luis Roldaao, and Dzmitry Tsishkou. Planerf: Svd unsupervised 3d plane regularization for nerf large-scale urban scene reconstruction. In *2024 International Conference on 3D Vision (3DV)*, pages 1291–1300. IEEE, 2024. 1

[34] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2161–2172, 2025. 2

[35] Lening Wang, Wenzhao Zheng, Dalong Du, Yunpeng Zhang, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, Jie Zhou, Jiwen Lu, et al. Stag-1: Towards realistic 4d driving simulation with video generation model. *arXiv preprint arXiv:2412.05280*, 2024. 1, 2

[36] Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. *arXiv preprint arXiv:2410.18079*, 2024. 1, 2

[37] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26024–26035, 2025. 1, 2, 4, 6, 7, 8, 3

[38] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21551–21561, 2024. 2

[39] Xiaogang Xu, Ying-Cong Chen, and Jiaya Jia. View independent generative adversarial network for novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7791–7800, 2019. 2

[40] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19595–19604, 2024. 2

[41] Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, et al. Streetcrafter: Street view synthesis with controllable video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 822–832, 2025. 1, 2, 7, 8, 3, 4

[42] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3, 4

[43] Chongjie Ye, Yinyu Nie, Jiahao Chang, Yuantao Chen, Yihao Zhi, and Xiaoguang Han. Gaustudio: A modular framework for 3d gaussian splatting and beyond. *arXiv preprint arXiv:2403.19632*, 2024. 2

[44] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 4

[45] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024. 3, 5, 7, 8, 1, 2, 4

[46] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5916–5926, 2025. 3

[47] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2

[48] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 1, 2

[49] Zhongrui Yu, Haoran Wang, Jinze Yang, Hanzhang Wang, Jiale Cao, Zhong Ji, and Mingming Sun. Sgd: Street view synthesis with gaussian splatting and diffusion prior. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3812–3822. IEEE, 2025. 2

[50] Xiaoyi Zeng, Kaiwen Song, Leyuan Yang, Bailin Deng, and Juyong Zhang. Oblique-merf: Revisiting and improving merf for oblique photography. In *International Conference on 3D Vision 2025*. 1

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[52] Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *arXiv preprint arXiv:2412.01718*, 2024. 1, 2

[53] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21336–21345, 2024. 4

[54] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 2

[55] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2