ArtiFixer: Enhancing and Extending 3D Reconstruction with Auto-Regressive Diffusion Models

RICCARDO DE LUTIO*, NVIDIA, USA
TOBIAS FISCHER, NVIDIA, Switzerland and ETHZ, Switzerland
YEN-YU CHANG, NVIDIA, USA and Cornell University, USA
YUXUAN ZHANG, NVIDIA, USA
JAY ZHANGJIE WU, NVIDIA, Canada
XUANCHI REN, NVIDIA, Canada, University of Toronto, Canada, and Vector Institute, Canada
TIANCHANG SHEN, NVIDIA, Canada, University of Toronto, Canada, and Vector Institute, Canada
KATARINA TOTHOVA, NVIDIA, Switzerland
ZAN GOJCIC, NVIDIA, Switzerland
HAITHEM TURKI*, NVIDIA, USA

Fig. 1. ArtiFixer enhances and extends existing 3D reconstructions in a highly efficient and scalable manner. Given an initial reconstruction, set of reference views, and an optional text prompt, it auto-regressively generates novel content that maintains a high degree of consistency with existing observations. ArtiFixer can directly produce hundreds of novel views in a single inference pass or serve as pseudo-supervision to improve the underlying 3D reconstruction. **Project page: https://research.nvidia.com/labs/sil/projects/artifixer**

Per-scene optimization methods such as 3D Gaussian Splatting provide state-of-the-art novel view synthesis quality but extrapolate poorly to under-observed areas. Methods that leverage generative priors to correct artifacts in these areas hold promise but currently suffer from two shortcomings. The first is scalability, as existing methods use image diffusion models or bidirectional video models that are limited in the number of views they can generate in a single pass (and thus require a costly iterative distillation process for consistency). The second is quality itself, as generators used in prior work tend to produce outputs that are inconsistent with existing scene content and fail entirely in completely unobserved regions. To solve these, we propose a two-stage pipeline that leverages two key insights. First, we train a powerful bidirectional generative model with a novel opacity mixing strategy that encourages consistency with existing observations while retaining the model's ability to extrapolate novel content in unseen areas. Second, we distill it into a causal auto-regressive model that generates hundreds of frames in a single pass. This model can directly produce novel views or serve as pseudo-supervision to improve the underlying 3D representation in a simple and highly efficient manner. We evaluate our method extensively and demonstrate that it can generate plausible reconstructions in scenarios where existing approaches fail completely. When measured on commonly benchmarked datasets, we outperform all existing baselines by a wide margin, exceeding prior state-of-the-art methods by 1-3 dB PSNR.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Rendering**.

Additional Key Words and Phrases: Image & Video Generative AI, Deep Image/Video Synthesis, Neural Rendering, Multi-View & 3D, Deep Learning, Machine Learning, Artificial Intelligence

*Equal contribution.

Authors' Contact Information: Riccardo de Lutio, NVIDIA, Santa Clara, USA, rdelutio@nvidia.com; Tobias Fischer, NVIDIA, Zurich, Switzerland and ETHZ, Zurich, Switzerland, tobiasf@nvidia.com; Yen-Yu Chang, NVIDIA, Santa Clara, USA and Cornell University, Ithaca, USA, yc2463@cornell.edu; Yuxuan Zhang, NVIDIA, New York, USA, alezhang@nvidia.com; Jay Zhangjie Wu, NVIDIA, Toronto, Canada, wjay@nvidia.com; Xuanchi Ren, NVIDIA, Toronto, Canada and University of Toronto, Toronto, Canada and Vector Institute, Toronto, Canada, xuanchir@nvidia.com; Tianchang Shen, NVIDIA, Toronto, Canada and University of Toronto, Toronto, Canada and Vector Institute, Toronto, Canada, frshen@nvidia.com; Katarina Tothova, NVIDIA, Zurich, Switzerland, ktothova@nvidia.com; Zan Gojcic, NVIDIA, Zurich, Switzerland, zgojcic@nvidia.com; Haithem Turki, NVIDIA, Seattle, USA, hturki@nvidia.com.

1 Introduction

High-quality novel view synthesis is essential for applications in virtual and augmented reality and closed-loop simulation for physical AI. These use cases require photorealistic rendering and the ability

to navigate complex environments under unconstrained camera motion. In recent years, two paradigms have emerged as dominant approaches to novel view synthesis: explicit 3D neural reconstruction [Kerbl et al. 2023; Mildenhall et al. 2020], and camera-controlled image or video generation [Ren et al. 2025; Zhou et al. 2025].

Neural reconstruction methods have matured significantly and now enable real-time rendering and high visual fidelity when trained from dense image collections with accurate camera poses. However, in the most widely used per-scene optimization setting, their performance remains fundamentally limited by the completeness and quality of the input observations. Regions that are sparsely observed or entirely missing during capture are poorly reconstructed, leading to artifacts, holes, or implausible geometry. While such deficiencies remain hidden near the training views, they are inevitably exposed during free navigation of the scene.

Conversely, recent video generative models have demonstrated the ability to synthesize photorealistic and temporally coherent content that is often indistinguishable from real-world videos [Google DeepMind 2024; NVIDIA et al. 2025; OpenAI 2024]. Despite this progress, precise camera control over extended sequences, long-term temporal consistency, and the accumulation of drift and hallucinations remain open challenges, limiting their applicability to interactive view synthesis.

Instead of treating reconstruction and generation as standalone alternatives, we aim to combine their complementary strengths: generative models serve as powerful priors to repair and complete imperfect reconstructions, while the explicit—albeit noisy and partial—3D representation provides a strong conditioning signal that grounds generation, mitigates long-term drift, and suppresses hallucinations. Recent methods have taken initial steps in this direction by training generative models to map degraded novel-view renderings to clean images and distilling the resulting improvements back into an underlying 3D representation [Fischer et al. 2025; Gao* et al. 2024; Wu et al. 2025d; Yu et al. 2024]. However, these approaches must navigate two fundamental trade-offs. First, they must balance temporal consistency and efficiency: some employ large bidirectional video generative models that provide strong temporal coherence but incur high computational cost [Fischer et al. 2025; Gao* et al. 2024; Wu et al. 2025b], while others rely on (multi-view) image-based generative models that are more efficient but limit temporal consistency and require progressive distillation strategies [Wu et al. 2025d, 2024]. Second, they face the trade-off between conditioning strength and generative capacity. Approaches [Wu et al. 2025b; Yu et al. 2024] that condition generation on corrupted renderings via concatenation or cross-attention risk altering the observed scene content, whereas methods [Fischer et al. 2025; Wu et al. 2025d] trained to directly map corrupted renderings to clean images are incapable of synthesizing missing content, due to the mode collapse in fully unobserved regions where all input pixels are black.

In our work, we follow this line of research by adapting a pre-trained bidirectional video diffusion model into a camera-controllable generator that maps corrupted renderings to clean images. To overcome the aforementioned limitations, we introduce two key contributions: **(i)** an opacity-aware noise mixing strategy that injects Gaussian noise into low-opacity regions, preventing mode collapse and preserving generative capacity in unobserved areas; and **(ii)** distillation of the bidirectional model into a few-step causal auto-regressive generator capable of producing arbitrarily long, temporally consistent videos while approaching the efficiency of prior image-based methods. In doing so, we demonstrate that even highly degraded 3D reconstructions provide sufficient conditioning signals to significantly simplify the distillation process. While recent work has begun incorporating explicit 3D representations as conditioning signals for auto-regressive video generation [Chen et al. 2025b; Wu et al. 2025c; Zhai et al. 2025], these approaches treat the 3D input as a fixed conditioning rather than an output to be improved. Our method closes this loop: the reconstruction conditions the generator, and the generator in turn enhances and extends the reconstruction, enabling both higher-quality video synthesis and improved 3D scene completeness. The resulting framework enables efficient improvement of the underlying 3D reconstruction and greatly outperforms a wide range of baselines across multiple benchmarks.

2 Related Work

Novel view synthesis from 3D representations. Neural Radiance Fields (NeRFs) [Mildenhall et al. 2020] and, more recently, 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023] have revolutionized the field of novel view synthesis by distilling sensor information (usually overlapping photos of a scene) into a 3D representation that can then be queried from arbitrary camera viewpoints. Because these representations are optimized on a per-scene basis, their ability to extrapolate beyond observed views is inherently limited, and they fail to render plausible content in sparsely observed or missing regions.

A large body of work seeks to mitigate these limitations through handcrafted geometric priors [Niemeyer et al. 2022; Somraj et al. 2023; Yang et al. 2023], pretrained depth [Deng et al. 2022; Roessle et al. 2022; Wang et al. 2023; Zhu et al. 2024] and normal [Yu et al. 2022] estimators, and adversarial networks [Roessle et al. 2023]. However, these approaches are sensitive to noise, difficult to balance with data terms, and yield only marginal improvements in denser captures. An alternative line of work trains feed-forward networks on large multi-scene datasets, which are used to enhance a scene-optimized NeRF/3DGS [Lu et al. 2025b; Zhou et al. 2023] or directly predict novel views [Chen et al. 2021; Lu et al. 2025a; Ren et al. 2024; Yu et al. 2021]. While these deterministic methods perform well near reference views, they often produce blurry results in ambiguous regions where the distribution of possible renderings is inherently multi-modal.

Diffusion models for novel view synthesis. An alternative strategy is to leverage the priors learned by generative diffusion models trained on internet-scale data to enhance novel view synthesis. Early works [Poole et al. 2023; Sargent et al. 2024; Wu et al. 2024] use a diffusion model as a learned critic during reconstruction optimization, but this incurs substantial computational overhead. More recent approaches [Fischer et al. 2025; Gao* et al. 2024; Liu et al. 2024, 2022; Wu et al. 2025d,b] directly generate multi-view–consistent images that can be consumed by a downstream 3D reconstruction pipeline. While this strategy improves training efficiency, it typically relies on iterative generation and distillation, in which new views are progressively distilled back into the 3D representation to satisfy

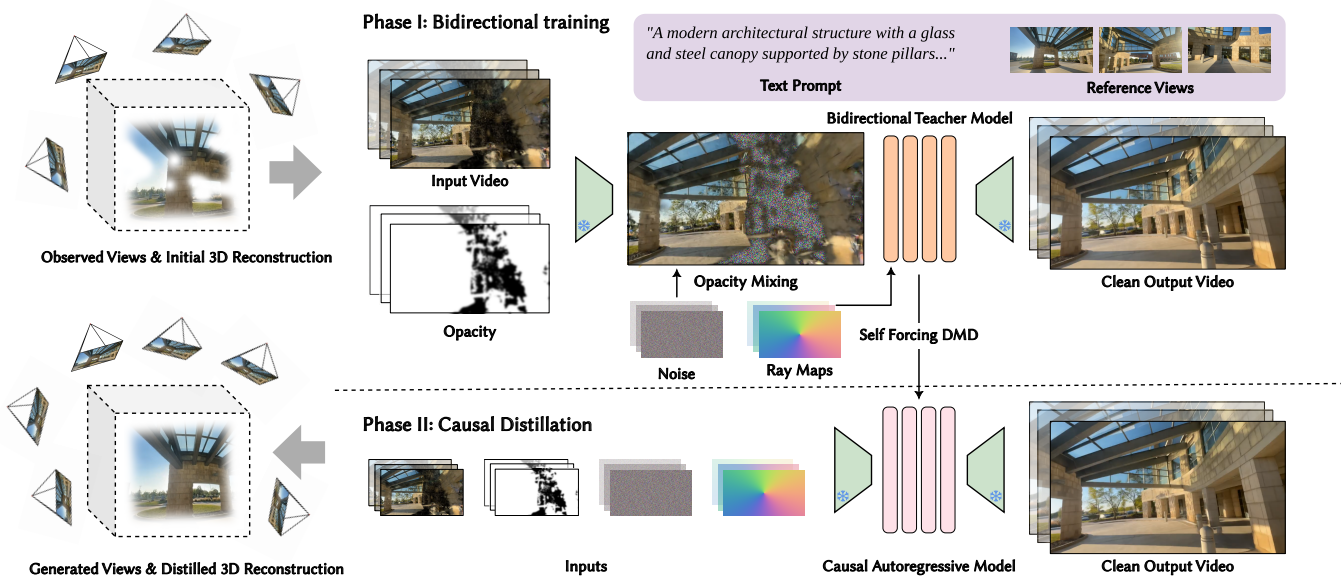

Fig. 2. **Method overview.** We first train a bidirectional flow matching model that transports degraded RGB renderings into clean outputs. We encode the input RGB into latent space and mix with Gaussian noise using the rendered opacity maps to avoid mode collapse in unseen regions. We inject fine-grained opacity information and camera control along with clean reference views and an optional text prompt. In the second phase of our pipeline, we distill the teacher into an auto-regressive causal model via Self Forcing-style DMD distillation [Huang et al. 2025], which can be directly used to render novel views or used as pseudo-supervision to distill back into the underlying 3D representation.

computational and consistency constraints. Lyra [Bahmani et al. 2026] sidesteps this iteration by distilling video diffusion knowledge into a feed-forward 3DGS generator, though it operates from a single image rather than enhancing an existing reconstruction. Building on the rapid progress of video generative models [Blattmann et al. 2023; Wan et al. 2025a], recent work reverses this paradigm. Rather than distilling generative outputs into a 3D representation, these methods treat the 3D representation as a conditioning signal for a generative model that directly synthesizes novel views [Kong et al. 2025; Ren et al. 2025]. Although this approach can improve the perceptual realism of novel views, it inherits limitations of the underlying generative models, including temporal inconsistencies, hallucinations, and imperfect camera control.

Auto-regressive video generation. While bidirectional video generation models synthesize all frames jointly, auto-regressive models generate frames sequentially using block-causal attention. Auto-regressive generation improves scalability and generation efficiency compared to bidirectional models, but often suffers from quality degradation over time, as each frame is conditioned on previously generated outputs, causing errors to accumulate [Yin et al. 2025b]. Several methods try to address the issue by better aligning the training scheme of these models with inference-time conditions, thereby reducing exposure bias [Cui et al. 2025; Huang et al. 2025; Liu et al. 2025]. A complementary line of research focuses on improving generation speed and controllability by exploiting temporal and spatial cues to select per-frame context [Kong et al. 2025; Li et al. 2025a; Shin et al. 2025; Wan et al. 2025b; Yang et al. 2025], enabling interactive auto-regressive world models [Hong et al. 2025]. Despite these advances, auto-regressive video models still lag behind explicit 3D representations in terms of spatial consistency, camera controllability, and rendering efficiency.

3 Preliminaries

3D Gaussian Splatting. 3DGS [Kerbl et al. 2023] represents a scene as a set of anisotropic 3D Gaussian primitives, each parameterized by a mean $\boldsymbol{\mu}_j$, covariance Σ_j, opacity σ_j, and view-dependent color $\mathbf{c}_j$. Novel views are rendered by projecting the primitives onto the target image plane and compositing in front-to-back depth order: $C(\mathbf{p}) = \sum_i \alpha_i \mathbf{c}_i \prod_{k<i}(1-\alpha_k)$, where α_i is the learned opacity scaled by the projected Gaussian evaluated at pixel $\mathbf{p}$. Primitive parameters are optimized per scene with a photometric reconstruction loss.

Video diffusion models. Diffusion models learn to transport samples between a data distribution $p_{data}(\mathbf{x})$ and a tractable prior, typically $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [Ho et al. 2020; Song et al. 2020]. Most video diffusion models [Blattmann et al. 2023] operate in a lower-dimensional latent space for computational efficiency. Flow matching [Lipman et al. 2023a; Liu et al. 2023], the framework used by our method, learns an ODE flow between two arbitrary endpoint distributions p_{src} and p_{tgt} by fitting a time-dependent vector field $\mathbf{v}_\theta(\mathbf{z}_t, t)$ whose induced probability path $\{p_t\}_{t\in[0,1]}$ satisfies $p_0 = p_{src}$ and $p_1 = p_{tgt}$. During training, we sample endpoint latents $\mathbf{z}_0 \sim p_{src}$ and $\mathbf{z}_1 \sim p_{tgt}$ and a time $t \in [0, 1]$, construct an intermediate latent via $\mathbf{z}_t \coloneqq (1-t)\mathbf{z}_0 + t\mathbf{z}_1$ with target velocity $\mathbf{v}_t \coloneqq \frac{d\mathbf{z}_t}{dt} = \mathbf{z}_1 - \mathbf{z}_0$, and fit the vector field using the conditional flow matching objective $\min_\theta \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left\|\mathbf{v}_\theta(\mathbf{z}_t, t) - \mathbf{v}_t\right\|_2^2$. At inference, we draw $\mathbf{z}_0 \sim p_{src}$ and numerically integrate the learned ODE from $t = 0$ to $t = 1$ to obtain $\mathbf{z}_1$ as a sample from p_{tgt}.

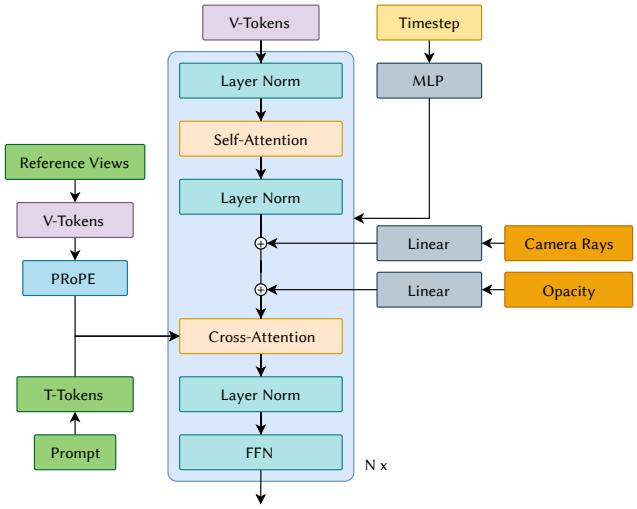

Fig. 3. **Transformer block.** We start from a pretrained text-to-video model [Wan et al. 2025a] and inject camera and opacity information into each transformer block via linear layers after applying self-attention and layer normalization. We patchify reference views into visual tokens, apply relative camera conditioning via PRoPE [Li et al. 2025b], and add K_n and V_n projections to the cross-attention operation. We zero-initialize f_r, f_o, and V_n to ensure compatibility with the pretrained initialization.

4 Method

Given an initial 3D reconstruction of a scene created from a sparse set of images, our goal is to generate artifact-free renderings from arbitrary camera viewpoints, including regions unobserved by input images, at interactive rates. Our solution is a controllable autoregressive video model that can either directly render arbitrary long novel-view renderings or provide pseudo-supervision to improve the underlying 3D reconstruction. We describe how to adapt a pretrained video diffusion model to serve as a bidirectional teacher in Sec. 4.1. We discuss causal distillation and the capabilities of the resulting model in Sec. 4.2. Fig. 2 illustrates our approach.

4.1 Bidirectional Training

Architecture. We start from a pretrained text-to-video model (Wan 2.1 T2V-14B [Wan et al. 2025a]), freeze its VAE and text encoder, and finetune the remaining components. Degraded renderings are encoded by the frozen VAE and 3D-patchified with $(t, h, w) = (1, 2, 2)$. We guide where to generate scene content through rendered opacity maps $\mathbf{O}$ and enable camera control in completely unobserved areas via Plücker raymaps $\mathbf{R}$. Both bypass the VAE entirely – we downscale their spatial dimensions to match the spatial compression factor of the VAE via the PixelUnshuffle operation [Paszke et al. 2019], encode them via per-block linear layers f_o and f_r (Fig. 3), and add the embeddings to the visual tokens:

$$T_r := T_s + f_r(\text{PixelUnshuffle}(\mathbf{R})) \tag{1}$$

$$T_o := T_r + f_o(\text{PixelUnshuffle}(\mathbf{O})), \tag{2}$$

where T_s denotes the token set after applying self-attention and layer-normalization. We found this strategy to be more computationally efficient than alternatives such as VAE encoding $\mathbf{R}$ and $\mathbf{O}$ while providing camera control even when the input rendering is entirely empty. To provide additional scene context, we encode clean reference views with the frozen VAE, patchified per-image along the batch dimension (no temporal compression). Each transformer block cross-attends from target tokens (Q) to concatenated reference tokens (K/V), with PRoPe [Li et al. 2025b] applied using target intrinsics/extrinsics for Q and reference intrinsics/extrinsics for K/V. f_r, f_o, and V_n are all zero-initialized to ensure compatibility with the pretrained initialization.

Opacity mixing. Most generative models start from Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ which is iteratively transformed into a latent video representation $\mathbf{z}$. Most prior work similarly starts from such noise, conditioning the generation process on the initial degraded rendering latent $\mathbf{z}_{deg}$ via channel-concatenation [Wu et al. 2025b; Yin et al. 2025a] or classifier-free guidance [Liu et al. 2022]. Although the resulting latent $\mathbf{z}_{enh}$ tends to be semantically similar to its degraded counterpart, notable inconsistencies remain, especially in high-artifact regions (Fig. 4). Several methods start directly from $\mathbf{z}_{deg}$ instead of noise [Fischer et al. 2025; Wu et al. 2025d], validating the insight that the source distribution should reflect what can already be rendered. While this encourages stronger consistency guarantees, it suffers from mode collapse in completely unseen areas: the source distribution collapses to a Dirac mass at zero in empty regions, hindering the ability to extrapolate high-quality renderings (Fig. 4). To address this, we mix Gaussian noise into low-opacity regions by downscaling $\mathbf{O}$ into $\mathbf{O}_z$ through max pooling to match $\mathbf{z}_{deg}$'s spatial dimensions (we retain fine-grained information via Eq. (2)) and deriving $\mathbf{z}_{mix} = \mathbf{O}_z \mathbf{z}_{deg} + (1 - \mathbf{O}_z)\boldsymbol{\epsilon}$ as the source distribution for our model. As no source information is lost from the max-pooling, this approach preserves the consistency benefits of starting from $\mathbf{z}_{deg}$ while gracefully interpolating to the standard Gaussian prior in entirely novel regions. This strategy is conceptually linked to inpainting methods [Avrahami et al. 2022; Kim et al. 2025; Mayet et al. 2025] that preserve known regions at low noise while pushing unknown regions toward the generative prior, though we operate with a continuous opacity signal rather than a binary mask. We formally derive compatibility with flow matching in Sec. A.

Data curation. Our goal is to not only correct artifacts in under-observed areas as in prior work [Fischer et al. 2025; Wu et al. 2025d] but also generate plausible content in entirely unseen areas. To do so, we generate paired reconstruction-ground truth samples from DL3DV-10K [Ling et al. 2024] with a camera selection strategy that encourages highly sparse reconstructions with large empty regions that the model must learn to inpaint. Given a set of camera poses with rotations $\mathbf{R}_i$ and translations $\mathbf{t}_i$, we first measure the camera pose distance function $d = ||\mathbf{R}_i - \mathbf{R}_j||_F + ||\mathbf{t}_i - \mathbf{t}_j||_2$, find the camera pair (P_1, P_2) with the largest distance, and seed groups G_1 and G_2. We assign the remaining cameras to G_1 or G_2 based on their distance to P_1 and P_2, and then sample 2-12 cameras with the largest inter-camera distance within each group to generate reconstructions of differing sparsity. We roughly align the camera scales of each reconstruction with a pretrained metric depth estimator [Wang et al. 2025] and prompt a vision-language model [Bai et al. 2025] for scene descriptions. We provide more details in Sec. G of the supplement.

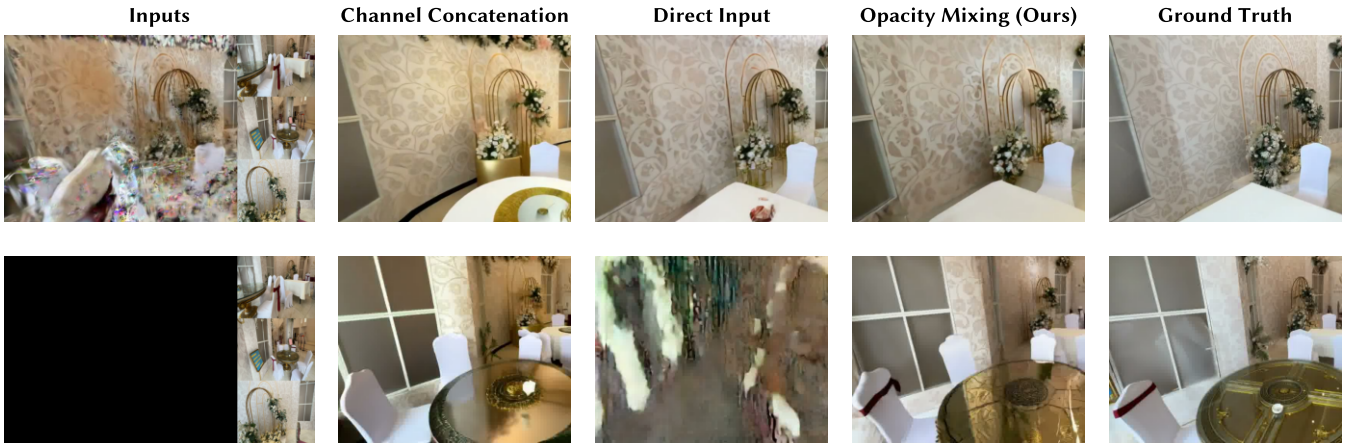

Fig. 4. **Opacity mixing.** Given a degraded rendering, set of reference views, and an optional text prompt (**left**), we predict an artifact-free rendering at a target viewpoint. Starting from Gaussian noise and channel concatenating the degraded rendering as in prior work [Wu et al. 2025b; Yin et al. 2025a] produces renderings that are semantically similar to the reference views, but with notable inconsistencies (such as the table in the **top row**). Directly starting from the degraded rendering instead of Gaussian noise improves consistency, but degrades quality noticeably when extrapolating to areas outside those covered by the degraded renderings (**bottom row**). Instead, we mix Gaussian noise into the rendering based on its opacity map. The resulting input retains the consistency benefits of the original while enabling a strong generative capability in entirely novel regions.

Optimization. Given an initial latent-encoded rendering $\mathbf{z}_{deg}$, which we transform into $\mathbf{z}_{mix}$, we train our model to predict its enhanced counterpart $\mathbf{z}_{enh}$ via conditional flow matching loss $\mathcal{L}_{cfm}$ [Lipman et al. 2023b]. We construct batches of paired reconstruction-ground truth data by sampling $N = 81$ frames along with the corresponding camera poses, text prompt (dropped with 10% probability), and a uniformly varying number of reference views (0-12). To enhance model's generative abilities and viewpoint controllability, we drop the last $K \leq N$ frames of the input (K is randomly chosen) by zeroing both the RGB rendering and opacity map while retaining the Plücker raymaps, so that the model must rebuild the ground truth from the prompt, reference views, and camera conditions alone.

4.2 Causal Distillation

Initialization. We initialize the causal model from the weights of the bidirectional teacher. To stabilize training, we follow a simpler strategy than the ODE initialization protocol of prior work [Huang et al. 2025; Shin et al. 2025; Yin et al. 2025b], which requires generating a dataset of ODE trajectories from the teacher model. Instead, we simply apply a block-causal mask, perturb each input frame with differing noise levels as in Diffusion Forcing [Chen et al. 2025a], and otherwise use the same inputs and training protocol as in Sec. 4.1.

Autoregressive rollout. After initialization, we adopt a training strategy similar to Self Forcing [Huang et al. 2025], where we generate video chunks sequentially and condition on previously generated chunks via KV caching, except that we continue applying dropout as in Sec. 4.1 as camera control and generation from pure noise otherwise degrade. We apply Distribution Matching Distillation (DMD) [Yin et al. 2024] to convert the model into a few-step generator ($N = 4$ in our experiments, although, outside of entirely novel regions, this can often be reduced to fewer steps with little noticeable difference as discussed in Sec. C of the supplement).

Long video generation. Existing methods rely on long-horizon training [Hong et al. 2025; Yang et al. 2025] to minimize error accumulation in long video rollouts. Although these strategies can be applied to our method, in practice we find our conditioning signals (notably the degraded rendering and reference views) sufficient to prevent error accumulation. We thus train with the same number of frames as in Sec. 4.1 and use a rolling KV cache during inference.

Although simple, this approach accelerates training convergence (due to training on a more diverse set of shorter videos for a given computational budget) and generalizes to arbitrary length videos, as shown in our experiments.

3D distillation. Prior work distills diffusion model outputs into 3D representations [Kerbl et al. 2023] for consistency purposes, as they otherwise exhibit temporal instability [Wu et al. 2025d] or are limited by number of frames bidirectional models can generate in a single pass [Fischer et al. 2025; Wu et al. 2025b]. As our auto-regressive model can sequentially generate arbitrary-length renderings, we are not limited by these constraints. However, 3D distillation is still sometimes desirable from an efficiency perspective, as these representations render orders of magnitude faster. To do so, existing methods require a progressive distillation process that alternates between view generation and 3D reconstruction, incurring significant training time overhead. In our case, as we can generate an arbitrary number of frames in a consistent manner, we adopt a more efficient approach by simply generating all desired novel views in a single pass before applying standard 3D reconstruction.

5 Experiments

We evaluate three variants of our method: ArtiFixer, which directly renders novel views from the auto-regressive generator, ArtiFixer 3D, which distills its outputs back into the underlying 3D representation, and ArtiFixer 3D+, which re-applies the auto-regressive

model as a post-processing step on top of ArtiFixer 3D (as in [Wu et al. 2025d]). We assess their ability to enhance in-the-wild captures against a wide range of prior work in Sec. 5.2 and their capacity to synthesize unobserved regions on a more challenging dataset split against a smaller set of relevant baselines in Sec. 5.3. We validate the contribution of individual components in Sec. 5.4.

5.1 Implementation

We implement our method in PyTorch [Paszke et al. 2019] and train it on 128 H100 GPUs, using a batch size of one per GPU (128 total). We use FlashAttention-3 [Shah et al. 2024] for acceleration. In our main experiments, we finetune the bidirectional model described in Sec. 4.1 for 15,000 iterations using AdamW [Loshchilov and Hutter 2019] with a learning rate of 1×10^{-5}. We then initialize the causal model for 5,000 iterations with the same learning rate, followed by 2,000 iterations of auto-regressive rollout and DMD training (≈15k GPU-hours total), using learning rates of 2×10^{-6} for the generator and 4×10^{-7} for the fake score function. For the ablations, we use a truncated schedule of 10,000 + 2,000 + 600 iterations on 64 GPUs to reduce computational cost (≈4k GPU-hours). We use 3DGUT [Wu et al. 2025a] with MCMC densification [Kheradmand et al. 2024] for the initial reconstructions used by our model. At test time, we use $K = 6$ uniformly sampled reference views for experiments matching the Difix3D+ protocol (Table 1) and all available input views otherwise (Tables 2 and 3). We use prompts generated by a vision-language model (Sec. G). Baselines are evaluated following their standard protocols.

5.2 Enhancing In-the-Wild Captures

Datasets. We run comparisons on the Nerfbusters dataset [Warburg et al. 2023] and DL3DV [Ling et al. 2024] using the splits provided by [Wu et al. 2025d] and on the Mip-NeRF 360 dataset [Barron et al. 2022] with the splits proposed by [Wu et al. 2024] and used in subsequent work [Gao* et al. 2024; Wu et al. 2025b].

Baselines. We compare ArtiFixer to an extensive set of baselines, including the original 3DGS [Kerbl et al. 2023] and 2DGS [Huang et al. 2024], NeRF variants [Barron et al. 2023; Tancik et al. 2023], non-generative sparse reconstruction methods [Li et al. 2024; Somraj et al. 2023; Yang et al. 2023; Zhu et al. 2024], and other diffusion-based work [Fischer et al. 2025; Gao* et al. 2024; Sargent et al. 2024; Warburg et al. 2023; Wu et al. 2025d, 2024, 2025b; Wynn and Turmukhambetov 2023; Yin et al. 2025a].

Metrics. We calculate PSNR, SSIM [Wang et al. 2004], LPIPS [Zhang et al. 2018], and FID score [Heusel et al. 2017] on Nerfbusters and DL3DV using the exact same protocol and metric implementations as Difix3D+ [Wu et al. 2025d]. On Mip-NeRF 360, we calculate PSNR, SSIM, and LPIPS across the 3, 6, and 9 view splits using the same implementations as GenFusion [Wu et al. 2025b].

Results. We present quantitative results for Nerfbusters and DL3DV in Table 1 and Mip-NeRF 360 in Table 2. We provide visual comparisons in Fig. 7 and Fig. 8. All ArtiFixer variants outperform all baselines by a substantial margin. Although the different variants produce similar renderings, ArtiFixer's are slightly sharper, while ArtiFixer 3D's are even more consistent with the source images

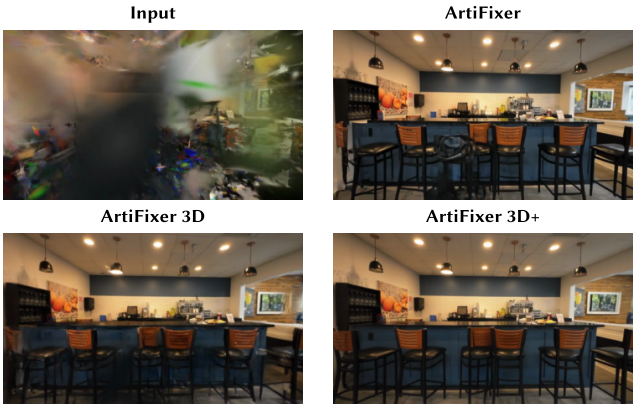

Fig. 5. **ArtiFixer variants.** Most visible differences occur in highly corrupted regions. ArtiFixer 3D's explicit 3D consistency improves fidelity with the source images and mitigates transient corruption (**middle**), at the cost of some sharpness, which ArtiFixer 3D+ restores. Nonetheless, all variants outperform prior work by a substantial margin.

at the cost of some blurriness due to its explicit 3D representation, leading to a minor increase in PSNR and SSIM and a small degradation in LPIPS and FID in Table 1. Re-applying the generator to the improved 3D reconstruction (ArtiFixer 3D+) restores some of this sharpness, leading to renderings that are crisper than ArtiFixer 3D and slightly more consistent than ArtiFixer (Fig. 5).

5.3 Novel Content Generation

Dataset. We evaluate novel content generation by following the sparse reconstruction protocol described in Sec. G on scenes from DL3DV, resulting in numerous "holes" that must be corrected in a manner consistent with existing observations.

Baselines. We compare to a smaller set of baselines most relevant to our work, notably 3DGUT [Wu et al. 2025a] as the base representation we provide as initial renderings to our method, image-based diffusion methods via Difix3D+ [Wu et al. 2025d] and Fixer [NVIDIA 2025], and approaches that build upon bidirectional video models [Ren et al. 2025; Wu et al. 2025b].

Results. We present quantitative results, using the same metrics as Table 1, in Table 3. We provide qualitative results in Fig. 6. All ArtiFixer variants outperform the next-best method (GenFusion [Wu et al. 2025b]) by almost 3 dB in PSNR. Gen3C [Ren et al. 2025] gives the next-best visually appealing results, but its conditioning often does not respect the source content, and its quality is upper-bounded by the depth estimator it uses to generate its 3D cache (in contrast to our purely data-driven approach). Difix3D+ [Wu et al. 2025d] and Fixer [NVIDIA 2025] generally fail to inpaint plausible context due to their deterministic conditioning.

5.4 Diagnostics

Ablations. We ablate the effectiveness of our opacity mixing strategy by comparing it to variants that instead use channel concatenation or omit the opacity mixing. We also measure the impact of the

Method	Nerfbusters [Warburg et al. 2023]				DL3DV [Ling et al. 2024]			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Nerfacto [Tancik et al. 2023]	17.29	0.621	0.402	134.65	17.16	0.581	0.430	112.30
3DGS [Kerbl et al. 2023]	17.66	0.678	0.327	113.84	17.18	0.588	0.384	107.23
Nerfbusters [Warburg et al. 2023]	17.72	0.647	0.352	116.83	17.45	0.606	0.370	96.61
GANeRF [Roessle et al. 2023]	17.42	0.611	0.354	115.60	17.54	0.610	0.342	81.44
NeRFLiX [Zhou et al. 2023]	17.91	0.656	0.346	113.59	17.56	0.610	0.359	80.65
Difix3D (Nerfacto) [Wu et al. 2025d]	18.08	0.653	0.328	63.77	17.80	0.596	0.327	50.79
Difix3D (3DGS) [Wu et al. 2025d]	18.14	0.682	0.287	51.34	17.80	0.598	0.314	50.45
Difix3D+ (Nerfacto) [Wu et al. 2025d]	18.32	0.662	0.279	49.44	17.82	0.613	0.283	41.77
Difix3D+ (3DGS) [Wu et al. 2025d]	18.51	0.686	0.264	41.77	17.99	0.602	0.293	40.86
ArtiFixer	19.83	0.701	0.254	37.78	19.73	0.672	0.231	20.85
ArtiFixer 3D	20.24	0.729	0.267	39.67	20.14	0.705	0.256	24.27
ArtiFixer 3D+	20.12	0.713	0.264	41.17	20.06	0.686	0.242	22.61

Table 1. **Artifact removal on Nerfbusters and DL3DV**. All ArtiFixer variants outperform prior methods by a considerable margin, improving PSNR by 2 dB.

Method	PSNR ↑			SSIM ↑			LPIPS ↓		
	3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view
Zip-NeRF [Barron et al. 2023]	12.77	13.61	14.30	0.271	0.284	0.312	0.705	0.663	0.633
3DGS [Kerbl et al. 2023]	13.06	14.96	16.79	0.251	0.355	0.447	0.576	0.505	0.446
2DGS [Huang et al. 2024]	13.07	15.02	16.67	0.243	0.338	0.423	0.580	0.506	0.449
FSGS [Zhu et al. 2024]	14.17	16.12	17.94	0.318	0.415	0.492	0.578	0.517	0.468
FreeNeRF [Yang et al. 2023]	12.87	13.35	14.59	0.260	0.283	0.319	0.715	0.717	0.695
SimpleNeRF [Somraj et al. 2023]	13.27	13.67	15.15	0.283	0.312	0.354	0.741	0.721	0.676
DiffusioNeRF [Wynn and Turmukhambetov 2023]	11.05	12.55	13.37	0.189	0.255	0.267	0.735	0.692	0.680
ZeroNVS [Sargent et al. 2024]	14.44	15.51	15.99	0.316	0.337	0.350	0.680	0.663	0.655
DNGaussian [Li et al. 2024]	14.00	15.21	16.72	0.301	0.356	0.397	0.620	0.604	0.603
FlowR [Fischer et al. 2025]	14.46	16.18	17.53	0.347	0.409	0.456	0.587	0.520	0.467
ReconFusion [Wu et al. 2024]	15.50	16.93	18.19	0.358	0.401	0.432	0.585	0.544	0.511
GenFusion [Wu et al. 2025b]	15.29	17.16	18.36	0.369	0.447	0.496	0.585	0.500	0.465
GSFixer [Yin et al. 2025a]	15.61	17.27	18.63	0.370	0.426	0.481	0.559	0.478	0.420
CAT3D [Gao* et al. 2024]	16.62	17.72	18.67	0.377	0.425	0.460	0.515	0.482	0.460
ArtiFixer	17.06	18.64	19.96	0.420	0.476	0.518	0.437	0.390	0.353
ArtiFixer 3D	17.29	18.95	20.24	0.451	0.526	0.598	0.440	0.382	0.327
ArtiFixer 3D+	17.51	18.95	20.16	0.444	0.498	0.537	0.441	0.396	0.359

Table 2. **Sparse view reconstruction methods on the Mip-NeRF360 dataset.** We exceed existing work by a wide margin across every metric.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
3DGUT [Wu et al. 2025a]	16.12	0.537	0.445	92.94
Difix3D (Nerfacto) [Wu et al. 2025d]	14.16	0.453	0.545	74.59
Difix3D (3DGS) [Wu et al. 2025d]	16.60	0.599	0.405	52.70
Difix3D+ (Nerfacto) [Wu et al. 2025d]	13.74	0.434	0.483	30.07
Difix3D+ (3DGS) [Wu et al. 2025d]	16.34	0.564	0.382	21.77
Fixer (offline) [NVIDIA 2025]	13.09	0.355	0.584	135.43
Fixer (online) [NVIDIA 2025]	13.93	0.443	0.535	79.44
Gen3C [Ren et al. 2025]	15.50	0.491	0.476	68.36
GenFusion [Wu et al. 2025b]	17.03	0.624	0.392	132.91
ArtiFixer	19.75	0.643	0.303	12.22
ArtiFixer 3D	19.92	0.673	0.306	16.28
ArtiFixer 3D+	20.15	0.662	0.307	13.91

Table 3. **Novel content generation.** We reconstruct DL3DV scenes following a protocol that creates large areas unobserved by training views. We outperform the next-best method (GenFusion) by almost 3 dB in PSNR.

Method	Direct Input	Opacity Mixing	Diffusion Forcing	PSNR↑	SSIM↑	LPIPS↓	FID↓
Channel Concatenation	✗	✗	✓	14.52	0.391	0.490	87.551
w/o Opacity Mixing	✓	✗	✓	17.34	0.440	0.429	87.058
w/o Initialization	✓	✓	✗	17.58	0.450	0.416	74.924
Full Method	✓	✓	✓	17.99	0.461	0.408	69.43

Table 4. **Diagnostics.** We evaluate reconstruction quality on Mip-NeRF 360. Denoising input renderings instead of conditioning via channel concatenation is crucial to producing outputs consistent with source images.

causal model weight initialization described in Sec. 4.2. We report results on Mip-NeRF 360 dataset averaged over all splits in Table 4 and show that our design choice of starting from the initial rendering instead of conditioning on it via channel concatenation is essential to rendering consistently with the source imagery. Our causal initialization method is not essential as the model still converges to a competitive level of quality, but provides a modest boost.

Model scale. To disentangle model scale from our other contributions, we train with Wan 2.1 T2V-1.3B and report results in Sec. D.

Timing. We report inference speed in Table 5 on a single GB300 GPU. Causal distillation with KV caching and few-step sampling yields a 70× speedup over the bidirectional Wan 2.1 14B and 1.3B backbones. With the 14B backbone, ArtiFixer and ArtiFixer 3D+

Method	FPS ↑
Wan 2.1 T2V-14B [Wan et al. 2025a]	0.12
Wan 2.1 T2V-1.3B [Wan et al. 2025a]	0.49
ArtiFixer/ArtiFixer 3D+ (14B)	8.36
ArtiFixer/ArtiFixer 3D+ (1.3B)	34.38
ArtiFixer 3D	268

Table 5. **Inference speed.** Causal distillation yields a 70× speedup over the bidirectional Wan 2.1 backbones. ArtiFixer 3D renders directly from 3DGUT. Additional configurations are reported in Table 7.

reach 8.36 FPS. Our 1.3B variant reaches 34.38 FPS. ArtiFixer 3D renders at native 3DGUT speed (268 FPS). Fewer denoising steps and context parallelism provide further gains (Sec. C).

6 Conclusion

Neural reconstruction and camera-controlled video generation provide complementary strengths for novel view synthesis. In this work, we introduced ArtiFixer, an auto-regressive video diffusion model that seeks to combine the advantages of both paradigms. ArtiFixer transforms corrupted renderings of reconstructed scenes into clean, temporally consistent frames, while retaining sufficient generative capacity to inpaint unobserved regions and the efficiency required for interactive use. The strong conditioning signal from the reconstructed scene significantly simplifies distillation and conversion to an auto-regressive formulation, enabling ArtiFixer to generate long video sequences with less quality degradation.

7 Acknowledgments

We thank Zian Wang and Nicholas Sharp for their helpful advice and feedback throughout this project.

References

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. 2025. MEt3R: Measuring Multi-View Consistency in Generated Images. In *CVPR*.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-Driven Editing of Natural Images. In *CVPR*.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. 2026. Lyra: Generative 3D Scene Reconstruction via Self-Distillation with Video Diffusion Models. In *ICLR*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL Technical Report. arXiv:2511.21631 [cs.CV] https://arxiv.org/abs/2511.21631

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *CVPR*.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2023. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. In *ICCV*.

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

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*. 14124–14133.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. 2025a. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NeurIPS* 37 (2025), 24081–24125.

Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. 2025b. FlexWorld: Progressively Expanding 3D Scenes for Flexiable-View Synthesis. In *NeurIPS*.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. 2025. Self-Forcing++: Towards Minute-Scale High-Quality Video Generation. *arXiv preprint arXiv:2510.02283* (2025).

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*. 12882–12891.

Tobias Fischer, Samuel Rota Bulò, Yung-Hsu Yang, Nikhil Keetha, Lorenzo Porzi, Norman Müller, Katja Schwarz, Jonathon Luiten, Marc Pollefeys, and Peter Kontschieder. 2025. FlowR: Flowing from Sparse to Dense 3D Reconstructions. In *ICCV*.

Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. 2024. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *NeurIPS*.

Google DeepMind. 2024. Veo: A Generative Model for High-Quality Video. https://deepmind.google/technologies/veo/. Accessed: 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS* 30 (2017).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* (2020).

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. 2025. RELIC: Interactive Video World Model with Long-Horizon Memory. arXiv:2512.04040 [cs.CV] https://arxiv.org/abs/2512.04040

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH Asia*.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. 2025. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. *NeurIPS*.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 2024. 3d gaussian splatting as markov chain monte carlo. *Advances in Neural Information Processing Systems* 37 (2024), 80965–80986.

Sora Kim, Sungho Suh, and Minsik Lee. 2025. RAD: Region-Aware Diffusion Models for Image Inpainting. In *CVPR*.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics* 36, 4 (2017).

Hanyang Kong, Xingyi Yang, Xiaoxu Zheng, and Xinchao Wang. 2025. WorldWarp: Propagating 3D Geometry with Asynchronous Video Diffusion. *arXiv preprint arXiv:2512.19678* (2025).

Vincent Leroy, Yohann Cabon, and Jerome Revaud. 2024. Grounding Image Matching innbsp;3D withnbsp;MASt3R. In *ECCV*.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. 2024. DNGaussian: Optimizing Sparse-View 3D Gaussian Radiance Fields with Global-Local Depth Normalization. In *CVPR*.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. 2025a. VMem: Consistent Interactive Video Scene Generation with Surfel-Indexed View Memory. In *ICCV*.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. 2025b. Cameras as Relative Positional Encoding. *NeurIPS*.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*. 22160–22169.

Yaron Lipman, {Ricky T.Q.} Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023a. Flow Matching for Generative Modeling. In *ICLR*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023b. Flow Matching for Generative Modeling. In *ICLR*.

Fangfu Liu, Wenqiang Wu, Hanyang Tan, Yueqi Yuan, Yikai Zhou, Junwei Liu, Kangjie Duan, Haowen Xie, Jingwen Pei, He Wang, et al. 2026. ReconX: Reconstruct Any Scene from Sparse Views with Video Diffusion Model. *IEEE Transactions on Image Processing* (2026).

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. 2025. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161* (2025).

Xinhang Liu, Jiaben Chen, Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. 2024. Deceptive-nerf: Enhancing nerf reconstruction using pseudo-observations from diffusion models. *ECCV*.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*.

Xi Liu, Chaoyi Zhou, and Siyu Huang. 2022. 3DGS-Enhancer: Enhancing Unbounded 3D Gaussian Splatting with View-consistent 2D Diffusion Priors. *NeurIPS*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. 2025a. InfiniCube: Unbounded and Controllable Dynamic 3D Driving Scene Generation with World-Guided Video Models. In *ICCV*.

Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiwei Li. 2025b. Matrix3D: Large Photogrammetry Model All-in-One. *CVPR* (2025).

Tsiry Mayet, Pourya Shamsolmoali, Simon Bernard, Eric Granger, Romain Hérault, and Clement Chatelain. 2025. TD-Paint: Faster Diffusion Inpainting Through Time Aware Pixel Conditioning. In *ICLR*.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *CVPR*.

NVIDIA. 2025. NVIDIA Fixer. https://huggingface.co/nvidia/Fixer. Accessed: 2026-01-26.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. 2025. Cosmos World Foundation Model Platform for Physical AI. https://arxiv.org/abs/2501.03575

OpenAI. 2024. Sora: Creating Video from Text. https://openai.com/sora. Accessed: 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*.

Xuanchi Ren, Yifan Lu, Hanxue Liang, Jay Zhangjie Wu, Huan Ling, Mike Chen, Francis Fidler, Sanja annd Williams, and Jiahui Huang. 2024. SCube: Instant Large-Scale Scene Reconstruction using VoxSplats. In *NeurIPS*.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *CVPR*.

Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. 2022. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*. 12892–12901.

Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. 2023. Ganerf: Leveraging discriminators to optimize neural radiance fields. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–14.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. 2024. ZeroNVS: Zero-Shot 360-Degree View Synthesis from a Single Image. In *CVPR*.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision. arXiv:2407.08608 [cs.LG] https://arxiv.org/abs/2407.08608

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. 2025. MotionStream: Real-Time Video Generation with Interactive Motion Controls. *arXiv preprint arXiv:2511.01266* (2025).

Nagabhushan Somraj, Adithyan Karanayil, and Rajiv Soundararajan. 2023. SimpleNeRF: Regularizing Sparse Input Neural Radiance Fields with Simpler Solutions. In *SIGGRAPH Asia*. doi:10.1145/3610548.3618188

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. 2023. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–12.

Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *ECCV*.

Haoran Wan, Jian Zhang, Richard Zhang, Jia-Bin Luo, Xinyang Fang, Lingbo Yang, Yanpei Cao, and Ying Shan. 2025b. Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation. *ACM Transactions on Graphics* (2025).

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025a. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. 2023. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *ICCV*. 9065–9076.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. 2025. MoGe-2: Accurate Monocular Geometry with Metric Scale and Sharp Details. In *CVPR*.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. 2023. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18120–18130.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. 2023. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. In *NeurIPS*.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. 2025d. DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models. In *CVPR*. 26024–26035.

Qi Wu, Janick Martinez Esturo, Ashkan Mirzaei, Nicolas Moenne-Loccoz, and Zan Gojcic. 2025a. 3DGUT: Enabling Distorted Cameras and Secondary Rays in Gaussian Splatting. In *CVPR*.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. 2024. ReconFusion: 3D Reconstruction with Diffusion Priors. In *CVPR*.

Sibo Wu, Congrong Xu, Binbin Huang, Geiger Andreas, and Anpei Chen. 2025b. GenFusion: Closing the Loop between Reconstruction and Generation via Videos. In *CVPR*.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. 2025c. Video world models with long-term spatial memory. In *NeurIPS*.

Jamie Wynn and Daniyar Turmukhambetov. 2023. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *CVPR*.

Jiawei Yang, Marco Pavone, and Yue Wang. 2023. FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization. In *CVPR*.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, and Song Hanand Yukang Chen. 2025. LongLive: Real-time Interactive Long Video Generation. (2025). arXiv:2509.22622 [cs.CV]

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. 2024. One-step Diffusion with Distribution Matching Distillation. In *CVPR*.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. 2025b. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. *CVPR*.

Xingyilang Yin, Qi Zhang, Jiahao Chang, Ying Feng, Qingnan Fan, Xi Yang, Chi-Man Pun, Huaqi Zhang, and Xiaodong Cun. 2025a. GSFixer: Improving 3D Gaussian Splatting with Reference-Guided Video Diffusion Priors. arXiv:2508.09667 [cs.CV] https://arxiv.org/abs/2508.09667

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *CVPR*.

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

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS* 35, 25018–25032.

Shangjin Zhai, Zhichao Ye, Jialin Liu, Weijian Xie, Jiaqi Hu, Zhen Peng, Hua Xue, Danpeng Chen, Xiaomeng Wang, Lei Yang, Nan Wang, Haomin Liu, and Guofeng Zhang. 2025. StarGen: A Spatiotemporal Autoregression Framework with Video Diffusion Model for Scalable and Controllable Scene Generation. In *CVPR*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. 2025. Stable Virtual Camera: Generative View Synthesis with Diffusion Models. *arXiv preprint* (2025).

Kun Zhou, Wenbo Li, Yi Wang, Tao Hu, Nianjuan Jiang, Xiaoguang Han, and Jiangbo Lu. 2023. NeRFLix: High-quality neural view synthesis by learning a degradation-driven inter-viewpoint mixer. In *CVPR*. 12363–12374.

Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. 2024. FSGS: Real-Time Few-Shot View Synthesis using Gaussian Splatting. In *ECCV*.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. 2026. FlashVSR: Towards Real-Time Diffusion-Based Streaming Video Super-Resolution. In *CVPR*.

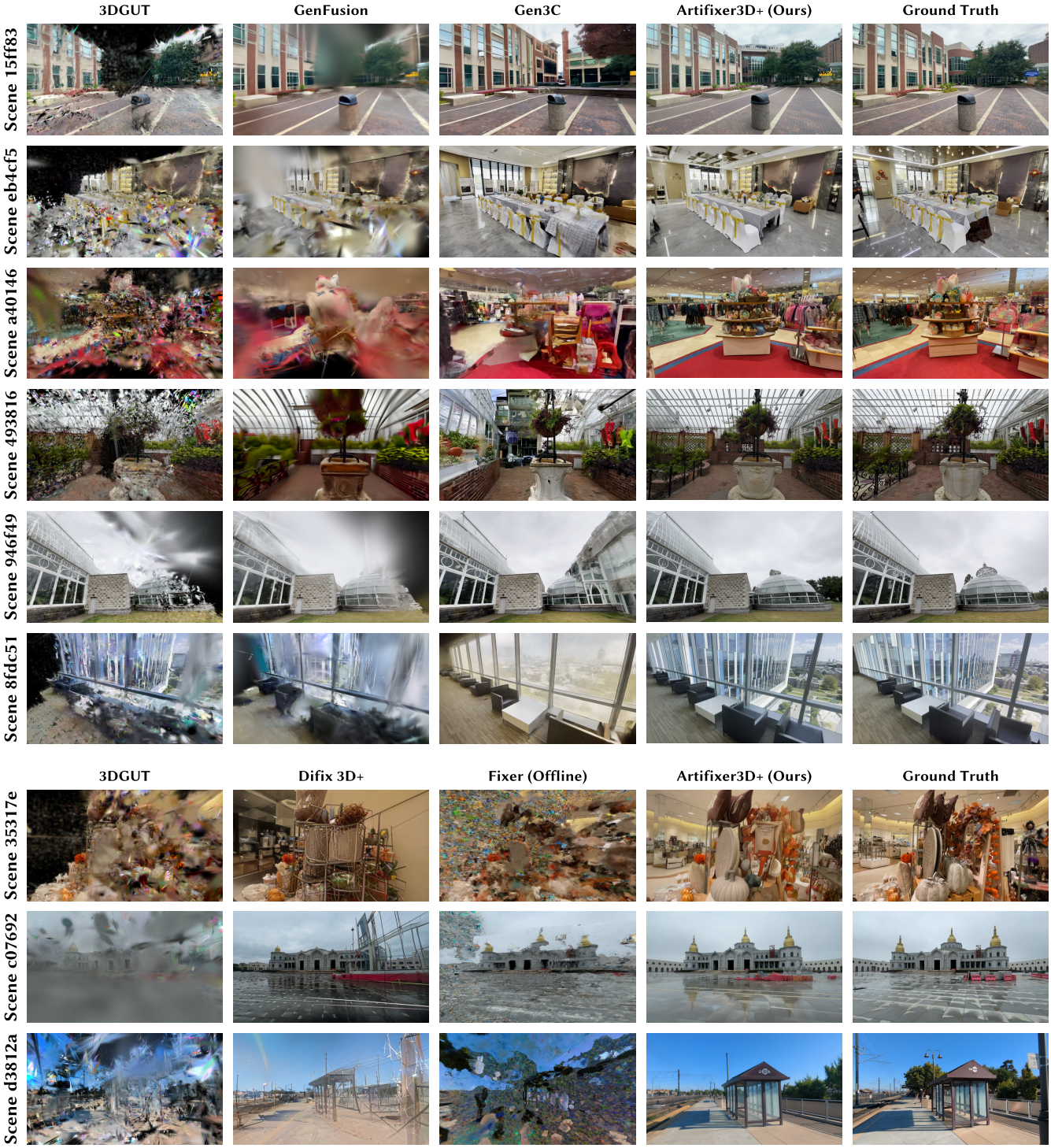

Fig. 6. **DL3DV results**. We compare ArtiFixer 3D+ to its initial 3DGUT [Wu et al. 2025a] input, two baselines that build upon bidirectional video diffusion models (**top rows**), and two that leverage image models (**bottom rows**). GenFusion [Wu et al. 2025b]'s video model generates 16 frames at a time, requiring an iterative distillation process that leads to blurry results, especially in empty areas. Gen3C [Ren et al. 2025]'s renderings are sharper but often do not respect the source content (background in **top row**), have incorrect geometry (**second row**), and exhibit color shift (**sixth row**). Methods that directly take renderings as input without opacity mixing [NVIDIA 2025; Wu et al. 2025d] fail to reconstruct empty regions. Our method can reconstruct plausible and consistent geometry even when the initial rendering is highly degraded. Please refer to our project website for comparison videos.

Fig. 7. **Mip-NeRF 360 results**. We present visualizations of Mip-NeRF's most challenging split (3-view). Our results far exceed all prior work both quantitatively and qualitatively. Our method is able to recover the correct geometry from the reference views even in scenarios where the input rendering is completely inaccurate (table in **third row**).

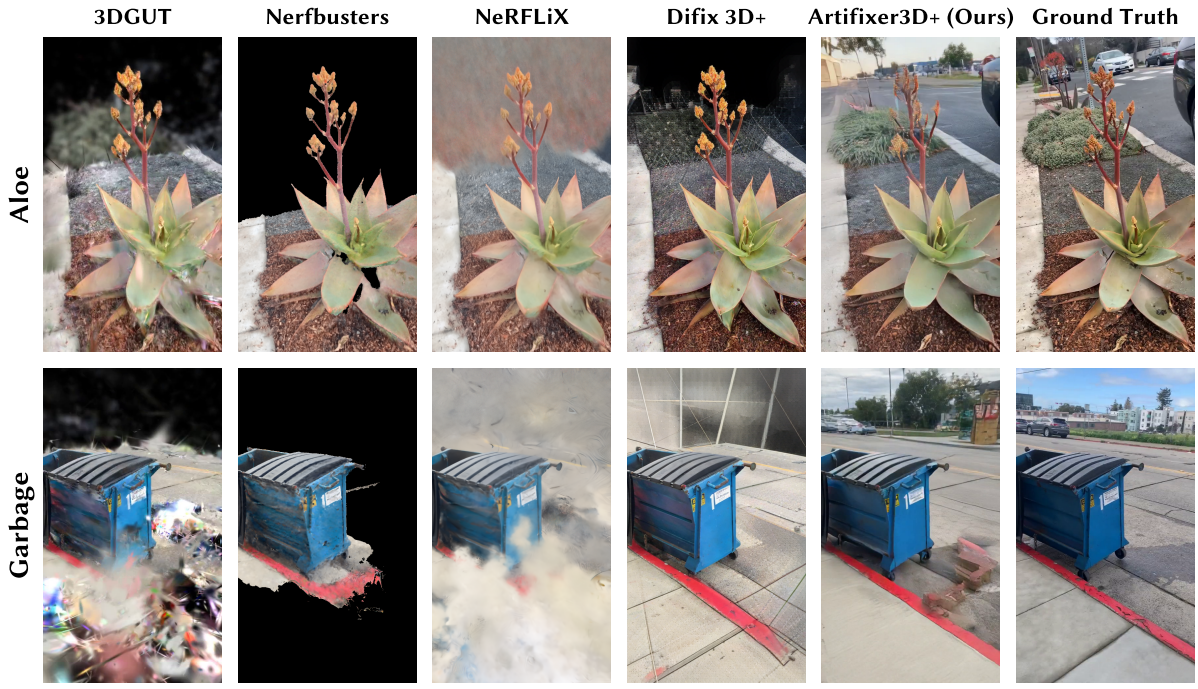

Fig. 8. **Nerfbusters results**. As with the other datasets, our method is the only one to generate plausible visuals in unseen regions while preserving the fidelity of the original content.

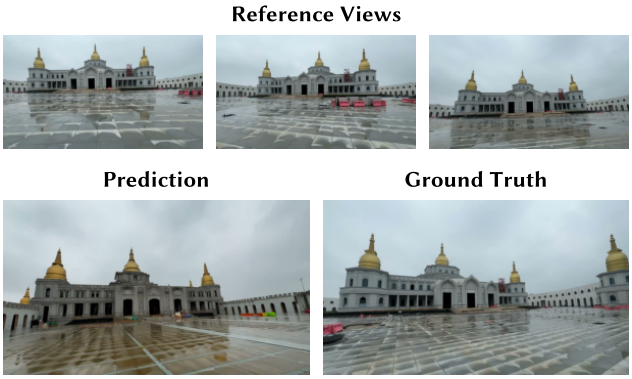

Fig. 9. **Reference views.** Without the initial rendering condition, ArtiFixer can generate predictions from the reference views. Although fidelity drops somewhat, the high-level structure of the scene remains intact.

A Opacity Mixing and Flow Matching

Our opacity mixing strategy is fully compatible with the conditional flow matching (CFM) framework [Lipman et al. 2023a] as the CFM loss $\mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left\|\mathbf{v}_\theta(\mathbf{z}_t, t, \text{cond}) - (\mathbf{z}_1 - \mathbf{z}_0)\right\|^2$ is valid for *any* joint distribution $q(\mathbf{z}_0, \mathbf{z}_1)$, not only $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In our setting, we define the source sample as:

$$\mathbf{z}_0 \coloneqq \mathbf{O}_z \mathbf{z}_{deg} + (1 - \mathbf{O}_z)\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where $\mathbf{O}_z$ is the spatially varying, downscaled opacity map and $\mathbf{z}_{deg}$ is the VAE-encoded degraded rendering. Let $\mathbf{z}_1$ denote the clean target latent. We sample a global scalar $t \sim \mathcal{U}[0, 1]$ and form the interpolant:

$$\mathbf{z}_t = (1 - t)\,\mathbf{z}_0 + t\,\mathbf{z}_1, \tag{4}$$

with target velocity $\mathbf{v}_t = \mathbf{z}_1 - \mathbf{z}_0$. The spatial variation introduced by $\mathbf{O}_z$ is encoded entirely in $\mathbf{z}_0$ and consequently propagates to both $\mathbf{z}_t$ and the target velocity $\mathbf{v}_t$, not to the scalar time variable t. No per-location timestep conditioning is required: the network receives $(\mathbf{z}_t, t, \text{cond})$ with a single global t, exactly as in standard flow matching.

At inference, we draw $\mathbf{z}_0 \sim q(\mathbf{z}_0)$ using the same opacity mixing procedure and integrate the learned ODE from $t = 0$ to $t = 1$ using the same global time parameterization.

B Conditioning

To probe which inputs drive output quality, we progressively strip conditioning signals. First, we drop the initial rendering, forcing the model to rely solely on reference views and camera rays. Although fidelity decreases, the model still recovers the high-level scene structure (Fig. 9). Next, we remove all conditioning except the text prompt, reverting to standard text-to-video generation; output quality remains comparable to the base Wan 2.1 model (Fig. 10).

We further quantify the contribution of text conditioning by comparing ArtiFixer 3D+ results with and without VLM-generated prompts in Table 6. Text conditioning provides a minor benefit in the most sparse settings (+0.14 dB PSNR on Mip-NeRF 360 with 3 views), but this effect diminishes with denser captures.

Fig. 10. **Text-to-video generation.** To illustrate our model's generative ability, we generate videos from text prompts alone. With opacity mixing, it retains similar quality to its base model [Wan et al. 2025a].

Dataset	ΔPSNR	ΔSSIM	ΔLPIPS
Mip-NeRF 360 (3 views)	+0.14	+0.003	−0.002
Mip-NeRF 360 (6 views)	+0.07	+0.002	−0.001
Mip-NeRF 360 (9 views)	+0.03	+0.003	−0.001
DL3DV	+0.02	0.000	−0.001
Nerfbusters	−0.07	+0.001	0.000

Table 6. **Text conditioning.** We measure the impact of VLM-generated prompts vs. no prompt for ArtiFixer 3D+. Text prompts provide a small benefit in sparse settings that diminishes with denser captures.

Method	GPUs	FPS ↑ 1 step	2 steps	3 steps	4 steps
ArtiFixer (14B)	1	29.42	16.07	11.03	8.36
ArtiFixer (14B)	4	58.72	35.91	24.65	19.18
ArtiFixer (1.3B)	1	86.75	57.76	43.20	34.38
ArtiFixer (1.3B)	4	101.77	69.44	53.77	49.24

Table 7. **Inference configurations.** Fewer denoising steps and context parallelism across multiple GPUs further improve throughput, with the 1.3B variant reaching up to 101.77 FPS.

C Denoising Steps

As ArtiFixer starts from renderings instead of pure noise, it is able to generate plausible visuals in fewer than four steps in most cases. Reducing the number of denoising steps significantly improves throughput, with context parallelism across multiple GPUs providing further gains (Table 7). However, sharpness and temporal consistency suffer somewhat in empty areas (Fig. 11). This is largely mitigated when revisiting previously explored areas in our ArtiFixer 3D and ArtiFixer 3D+ variants, as the 3D distillation process provides strong conditioning for subsequent generations.

D Additional Experiments

Model scale. To disentangle the contribution of our method from backbone capacity, we train the full pipeline with Wan 2.1 T2V-1.3B and report ArtiFixer 3D+ results in Tables 8 and 9. For reference, GenFusion [Wu et al. 2025b] uses a 1.4B-parameter backbone, GS-Fixer [Yin et al. 2025a] 5B, and Gen3C [Ren et al. 2025] 7B. Our 1.3B variant matches CAT3D [Gao* et al. 2024] within 0.02 dB on the 3-view Mip-NeRF 360 split and exceeds all other baselines.

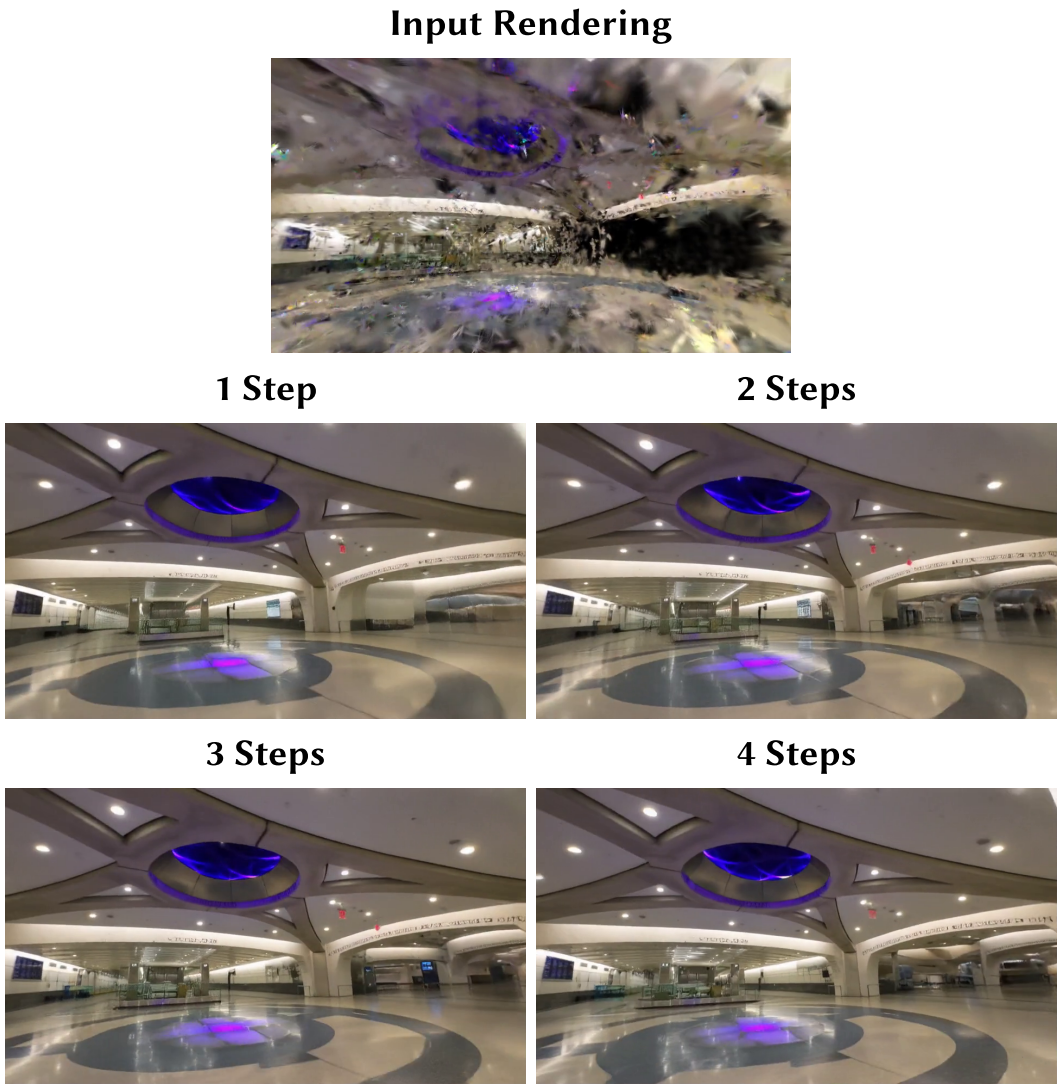

Fig. 11. **Denoising steps**. We vary the number of denoising steps when beginning from the initial degraded rendering. ArtiFixer can render plausible content in as few as 1 step, although sharpness and temporal consistency suffer somewhat in empty areas.

Method	PSNR ↑			SSIM ↑			LPIPS ↓		
	3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view
GenFusion [Wu et al. 2025b]	15.29	17.16	18.36	0.369	0.447	0.496	0.585	0.500	0.465
GSFixer [Yin et al. 2025a]	15.61	17.27	18.63	0.370	0.426	0.481	0.559	0.478	0.420
CAT3D [Gao* et al. 2024]	16.62	17.72	18.67	0.377	0.425	0.460	0.515	0.482	0.460
ArtiFixer 3D+ (1.3B)	16.60	18.04	19.44	0.414	0.466	0.513	0.486	0.435	0.394
ArtiFixer 3D+ (14B)	17.51	18.95	20.16	0.444	0.498	0.537	0.441	0.396	0.359

Table 8. **Impact of model scale on Mip-NeRF 360.** Our 1.3B variant matches CAT3D within 0.02 dB on the 3-view split and exceeds other video model baselines despite using fewer parameters.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
GenFusion [Wu et al. 2025b]	17.03	0.624	0.392	132.91
Gen3C [Ren et al. 2025]	15.50	0.491	0.476	68.36
ArtiFixer 3D+ (1.3B)	19.04	0.635	0.352	22.3
ArtiFixer 3D+ (14B)	20.15	0.662	0.307	13.91

Table 9. **Impact of model scale on novel content generation (DL3DV).** Even with a 1.3B backbone, ArtiFixer 3D+ outperforms the other video model baselines by a wide margin.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
3DGS [Kerbl et al. 2023]	9.57	0.108	0.779
SparseNeRF [Wang et al. 2023]	9.23	0.191	0.632
DNGaussian [Li et al. 2024]	10.23	0.156	0.643
ReconX [Liu et al. 2026]	14.28	0.394	0.564
ArtiFixer 3D+	14.75	0.464	0.463

Table 10. **Tanks and Temples (2-view).** ArtiFixer 3D+ outperforms all baselines.

Method	MASt3R ↓	RAFT ↓
Fixer [NVIDIA 2025]	0.1288	0.1236
Difix3D+ [Wu et al. 2025d]	0.0974	0.0959
GenFusion [Wu et al. 2025b]	0.0817	0.0786
Gen3C [Ren et al. 2025]	0.0766	0.0757
ArtiFixer	0.0749	0.0749
ArtiFixer 3D+	0.0697	0.0697
ArtiFixer 3D	0.0646	0.0647

Table 11. **Multi-view consistency.** We measure multi-view consistency via MEt3R [Asim et al. 2025] with MASt3R and RAFT backbones. All ArtiFixer variants outperform baselines, with ArtiFixer 3D achieving the best results due to its explicit multi-view-consistent 3D representation.

Tanks and Temples. To further evaluate generalization, we report results on the Tanks and Temples dataset [Knapitsch et al. 2017] using the 2-view setting from ReconX [Liu et al. 2026] in Table 10.

Multi-view consistency. We evaluate multi-view consistency using MEt3R [Asim et al. 2025] with MASt3R [Leroy et al. 2024] depth-based reprojection and RAFT [Teed and Deng 2020] optical flow-based warping backbones in Table 11. All ArtiFixer variants outperform baselines, with ArtiFixer 3D achieving the best consistency due to its explicit 3D representation.

E Limitations

While ArtiFixer reaches interactive rates, it remains significantly slower than direct rendering from neural scene representations. Decoding in temporal chunks also introduces latency that may be undesirable for applications such as embodied AI. Additionally, the ArtiFixer and ArtiFixer 3D+ variants are limited to 720p by the backbone video model, whereas ArtiFixer 3D renders at the native resolution of the underlying 3D representation. As with other video diffusion models, our method can occasionally blur fine details and text, and may introduce subtle color shifts when the rendering condition is absent or highly degraded. Promising directions for future work include further reducing denoising steps, enabling single-frame decoding while maintaining temporal coherence, and applying video super-resolution [Zhuang et al. 2026] to close the resolution gap.

F Societal Impact

ArtiFixer synthesizes photorealistic scene content and can plausibly inpaint unobserved regions, raising concerns about potential misuse for generating deceptive visual media. Appropriate safeguards such as watermarking generated content [Wen et al. 2023] should be considered for deployment. From an environmental perspective, training our 14B-parameter model requires approximately 15k GPU-hours on H100 hardware. Our truncated training schedule achieves near-full quality at roughly 25% of this cost, and our 1.3B-parameter variant further reduces training compute while remaining competitive with prior work.

G Sparse Reconstruction

Camera Sampling. We describe our camera sampling strategy in Algorithm 1. Given a set of camera poses $\mathbf{P}$, we define the pairwise distance between two poses as $d = ||\mathbf{R}_i - \mathbf{R}_j||_F + ||\mathbf{t}_i - \mathbf{t}_j||_2$. We initialize the clustering process by identifying the pair (P_1, P_2) that maximizes this distance and using them as seeds for groups G_1 and G_2. The remaining cameras are assigned to the group of their nearest seed. Finally, to evaluate varying levels of sparsity, we apply farthest point sampling within each group to select subsets of size $K = \{2, \cdots, 12\}$.

ALGORITHM 1: CameraSampling

Input: Camera poses $\mathbf{P}$, Selection count K, Distance function d
Output: Selected subsets $\mathcal{S}_1 \subset G_1$ and $\mathcal{S}_2 \subset G_2$

```
/* 1. Find global farthest camera pair                    */
```
$(P_1, P_2) \leftarrow \text{argmax}_{P_i, P_j \in \mathbf{P}} d(P_i, P_j)$;
```
/* 2. Cluster: Assign cameras to nearest seed camera      */
```
$G_1 \leftarrow \{P \in \mathbf{P} \mid D(P, P_1) \leq D(P, P_2)\}$;
$G_2 \leftarrow \mathbf{P} \setminus G_1$;
```
/* 3. Select Top-K points in EACH group                   */
```
foreach $i \in \{1, 2\}$ **do**
    $\mathcal{S}_i \leftarrow \{P_i\}$; // Start with the seed camera
    while $|\mathcal{S}_i| < K$ *and* $|\mathcal{S}_i| < |G_i|$ **do**
        /* Find pose maximizing distance to current
           selection   */
        $P_{next} \leftarrow \text{argmax}_{P \in G_i \setminus \mathcal{S}_i} \left(\min_{s \in \mathcal{S}_i} D(P, s)\right)$;
        $\mathcal{S}_i \leftarrow \mathcal{S}_i \cup \{P_{next}\}$;
    end
end
return $\mathcal{S}_1, \mathcal{S}_2$

Reconstruction. We generate the initial reconstructions we pass to the ArtiFixer model using the official 3DGUT implementation [Wu et al. 2025a] with MCMC [Kheradmand et al. 2024] sampling (reconstructions used during training are prepared offline). We run each reconstruction for 10,000 iterations, taking slightly less than 10 minutes per reconstruction.

Captioning. We generate captions for each DL3DV scene from Qwen3-VL-30B-A3B-Instruct [Bai et al. 2025] on different frame subsets to encourage prompt diversity. Similar to [Hong et al. 2025], we suppress descriptions of ego-camera movement to avoid entanglement with camera ray conditioning. We use the prompt below:

You are a video captioning specialist whose goal is to generate high-quality English prompts by referring to the details of the user's input videos. Your task is to carefully analyze the content, context, and actions within the video, and produce a complete, expressive, and natural-sounding caption that accurately conveys the scene. The caption should preserve the original intent and meaning of the video while enhancing its clarity and descriptive richness. Strictly adhere to the formatting of the examples provided.

Task Requirements: 1. You need to describe the main subject of the video in detail, including their appearance, actions, expressions, and the surrounding environment. 2. You should never describe any details about the camera movement or camera angles. 3. Your output should convey natural movement attributes, incorporating natural actions related to the described subject category, using simple and direct verbs as much as possible. 4. You should reference the detailed information in the video, such as character actions, clothing, backgrounds, and emphasize the details in the photo. 5. Control the output prompt to around 80-100 words. 6. No matter what language the user inputs, you must always output in English.

Example of the English prompt: 1. A Japanese fresh film-style photo of a young East Asian girl with double braids sitting by the boat. The girl wears a white square collar puff sleeve dress, decorated with pleats and buttons. She has fair skin, delicate features, and slightly melancholic eyes, staring directly at the camera. Her hair falls naturally, with bangs covering part of her forehead. She rests her hands on the boat, appearing natural and relaxed. The background features a blurred outdoor scene, with hints of blue sky, mountains, and some dry plants. The photo has a vintage film texture. A medium shot of a seated portrait. 2. An anime illustration in vibrant thick painting style of a white girl with cat ears holding a folder, showing a slightly dissatisfied expression. She has long dark purple hair and red eyes, wearing a dark gray skirt and a light gray top with a white waist tie and a name tag in bold Chinese characters. The background has a light yellow indoor tone, with faint outlines of some furniture visible. A pink halo hovers above her head, in a smooth Japanese cel-shading style. A close-up shot from a slightly elevated perspective. 3. CG game concept digital art featuring a huge crocodile with its mouth wide open, with trees and thorns growing on its back. The crocodile's skin is rough and grayish-white, resembling stone or wood texture. Its back is lush with trees, shrubs, and thorny protrusions. With its mouth agape, the crocodile reveals a pink tongue and sharp teeth. The background features a dusk sky with some distant trees, giving the overall scene a dark and cold atmosphere. A close-up from a low angle. 4. In the style of an American drama promotional poster, Walter White sits in a metal folding chair wearing a yellow protective suit, with the words "Breaking Bad" written in sans-serif English above him, surrounded by piles of dollar bills and blue plastic storage boxes. He wears glasses, staring forward, dressed in a yellow jumpsuit, with his hands resting on his knees, exuding a calm and confident demeanor. The background shows an abandoned, dim factory with light filtering through the windows. There's a noticeable grainy texture. A medium shot with a straight-on close-up of the character.

Directly output the English text.