Novel View Extrapolation with Video Diffusion Priors

Kunhao Liu[1] Ling Shao[2] Shijian Lu[1]

[1]Nanyang Technological University [2]UCAS-Terminus AI Lab, UCAS

Figure 1. We introduce **ViewExtrapolator**, a novel approach that leverages the generative priors of Stable Video Diffusion for novel view extrapolation, where the novel views lie far beyond the range of the training views. ViewExtrapolator effectively refines the artifact-prone renderings (left side of arrows) of radiance fields or point clouds, to more realistic renderings with fewer artifacts (right side of arrows).

Abstract

The field of novel view synthesis has made significant strides thanks to the development of radiance field methods. However, most radiance field techniques are far better at novel view interpolation than novel view extrapolation where the synthesis novel views are far beyond the observed training views. We design ViewExtrapolator, a novel view synthesis approach that leverages the generative priors of Stable Video Diffusion (SVD) for realistic novel view extrapolation. By redesigning the SVD denoising process, ViewExtrapolator refines the artifact-prone views rendered by radiance fields, greatly enhancing the clarity and realism of the synthesized novel views. ViewExtrapolator is a generic novel view extrapolator that can work with different types of 3D rendering such as views rendered from point clouds when only a single view or monocular video is available. Additionally, ViewExtrapolator requires no fine-tuning of SVD, making it both data-efficient and computation-efficient. Extensive experiments demonstrate the superiority of ViewExtrapolator in novel view extrapolation. Project page: https://kunhao-liu.github.io/ViewExtrapolator/.

1. Introduction

The field of novel view synthesis has witnessed remarkable advancements, largely driven by the development of radiance field methods such as NeRF [32], Instant-NGP [33], 3D Gaussian Splatting [20], etc. These methods have revolutionized the way we render photorealistic images of novel

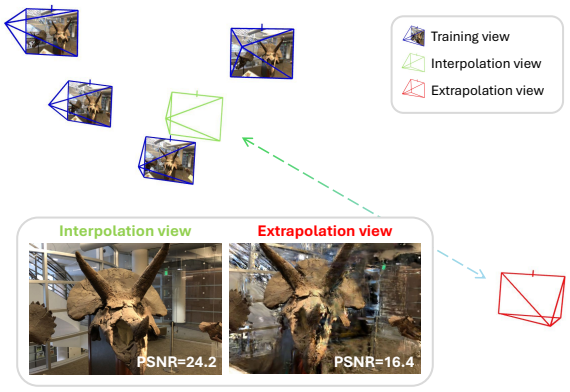

Figure 2. The setting differences between novel view *interpolation* and novel view *extrapolation*: Radiance fields excel at novel view interpolation but struggle at novel view extrapolation.

views by learning continuous volumetric scene representations from a set of training views.

The success of radiance fields is especially notable in novel view *interpolation* when the synthesized novel view lies within or near the convex hull of the training views. For the case of novel view *extrapolation* where the novel views move significantly beyond the range of training views, most existing radiance field methods struggle due to the lack of observed training data around the novel views [41]. However, novel view extrapolation is crucial for delivering an immersive 3D experience, allowing users to explore reconstructed radiance fields freely beyond the initial training views. Fig. 2 illustrates the setup differences between novel view interpolation and novel view extrapolation, as well as how they affect the synthesized novel views.

We design ViewExtrapolator, a novel view extrapolation technique that introduces the generative priors of Stable Video Diffusion (SVD) [4] for generating realistic extrapolative novel views. Given a reconstructed radiance field from training views with limited range, ViewExtrapolator first renders a video that starts from a training view and gradually transits to a distant extrapolative novel view. While the early video frames exhibit high-quality renderings, artifacts gradually arise in the ensuing video frames when the view goes beyond the training views. The artifacts become especially obvious around the extrapolated regions due to the lack of observed data in training. We introduce SVD as trained over large-scale natural videos to refine the artifact-prone novel-view frames. Specifically, we redesign the denoising process to guide SVD to preserve the original scene content by modifying the ODE derivative toward the artifact-prone videos. In addition, we design guidance annealing and resampling annealing that reduce the influence of the artifacts in the denoising steps and resampling steps [29], respectively, inpainting unseen regions and refining the visual quality throughout the denoising process effectively.

ViewExtrapolator has two unique features in novel view extrapolation. First, it is generic and can work with different 3D rendering approaches with little adaptation. For example, it can be directly applied to 3D renderings by point clouds as derived by depth estimation from a single view or monocular video. Second, ViewExtrapolator is an inference-stage method that does not require fine-tuning the SVD model. This makes it both data-efficient and computation-efficient, paving the way for more applicable and accessible novel view extrapolation.

The contributions of this work can be summarized in three key aspects. *First*, we introduce ViewExtrapolator, a novel training-free pipeline that leverages the generative priors of SVD for novel view extrapolation. *Second*, we design guidance annealing and resampling annealing that eliminate artifacts and enable high-quality inpainting of unseen regions, enhancing the visual fidelity of the rendered novel views effectively. *Third*, extensive experiments over various 3D rendering approaches demonstrate the superiority and broad applicability of ViewExtrapolator in novel view extrapolation.

2. Related Work

Radiance fields. Radiance fields [32] have emerged as a powerful representation of 3D scenes, driving advancements in novel view synthesis. They model 3D space by mapping radiance and density to arbitrary 3D coordinates, where pixel colors are rendered by aggregating the radiance values of sampled 3D points through volume rendering [30]. Radiance fields can be implemented using various methods, including MLPs [1, 2, 32, 55], decomposed tensors [5, 7, 10, 23, 24], hash tables [33], voxels [9, 42], and 3D Gaussians [20, 25, 28]. Numerous studies have been proposed to enhance the view synthesis process. For instance, Mip-NeRF [1, 2] improves rendering quality using anti-aliased conical frustums. Instant-NGP [33] accelerates training speed by modeling 3D volumes with multi-resolution hash tables. 3D Gaussian Splatting [20] achieves real-time rendering through rasterization with explicitly parameterized 3D Gaussians. However, these approaches generally require dense scene observations and lack the generative capacity for extrapolating beyond observed views, limiting their effectiveness in novel view extrapolation. While methods like ExtraNeRF [41] and RapNeRF [54] attempt to address novel view extrapolation, ExtraNeRF's extrapolation range is limited, and RapNeRF is restricted to object-level view synthesis. In contrast, ViewExtrapolator can render scene-level realistic novel views that lie far beyond the range of the training views.

Diffusion priors for view synthesis. Recent work has explored the generative priors of diffusion models [14] for

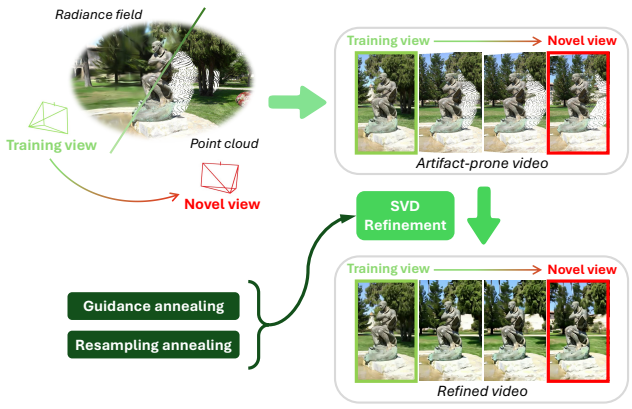

Figure 3. **Overview of the proposed ViewExtrapolator.** We render an artifact-prone video from the closest training view to an extrapolative novel view, and then refine it by guiding SVD to preserve the original scene content and eliminate the artifacts with guidance annealing and resampling annealing.

novel view synthesis. Early efforts focused on distilling the knowledge of 2D text-to-image diffusion models [36] into 3D using Score Distillation Sampling [35, 45], synthesizing 3D objects from text and images [18, 22, 43, 47]. Several studies fine-tune or train 2D diffusion models on multi-view or camera-pose-conditioned datasets to strengthen 3D priors [6, 12, 15, 26, 26, 27, 37, 39, 40, 46, 50, 51], though most of them focus on object-level synthesis. For scene-level synthesis, approaches like ExtraNeRF [41], DiffusioNeRF [52], and Nerfbusters [49] incorporate geometry-informed diffusion models for improved scene-level 3D reconstruction, while methods like Zero-NVS [37], Reconfusion [51], and CAT3D [11] employ diffusion models trained on large-scale multi-view datasets to enable scene-level few-shot reconstruction. In addition, MotionCtrl [48], CameraCtrl [13], ViVid-1-to-3 [21], and SV3D [44] leverage video diffusion models fine-tuned on camera trajectories for view synthesis, whereas NVS-solver [53] and CamTrol [16] utilize a training-free approach for camera control. Different from these developments, we propose a training-free approach for novel view extrapolation with video diffusion priors, paving a more applicable and accessible way in novel view synthesis.

3. Method

We tackle the challenges of novel view extrapolation by leveraging the generative priors of a large-scale video diffusion model SVD (Sec. 3.1) for refining artifact-prone videos as rendered by radiance fields or point clouds (Sec. 3.2). Specifically, we guide the SVD model to preserve the original scene content by modifying the ODE derivative towards the artifact-prone videos (Sec. 3.3). Additionally, we design guidance annealing and resampling annealing, which enable SVD to effectively refine the artifact-prone videos during the denoising process (Sec. 3.4). Fig. 3 illustrates the overview of the proposed ViewExtrapolator.

3.1. Preliminaries on Stable Video Diffusion

SVD [4] is an image-to-video diffusion model that conditions on an input image. By default, it generates a natural video that starts with the conditional image and autonomously evolves with camera movements and scene dynamics. As a diffusion model [14], SVD produces the video by progressively denoising a Gaussian noise. Given the noisy video latent $\mathbf{x}_t$ and the noise level σ_t at the diffusion time step $t \in [1, T]$, SVD parameterizes the denoising process following the EDM pre-conditioning framework [19]:

$$\hat{\mathbf{x}}_0 = c_{\text{skip}}(\sigma_t)\mathbf{x}_t + c_{\text{out}}(\sigma_t)F_{\boldsymbol{\theta}}(c_{\text{in}}(\sigma_t)\mathbf{x}_t; c_{\text{noise}}(\sigma_t)), \quad (1)$$

where $\hat{\mathbf{x}}_0$ is the predicted clean video at the current time step t, $c_{\text{skip}}, c_{\text{out}}, c_{\text{in}}$, and c_{noise} denote the predefined pre-conditioning functions, and $F_{\boldsymbol{\theta}}$ is the trainable network with parameters $\boldsymbol{\theta}$. With the current predicted clean video $\hat{\mathbf{x}}_0$, the ODE derivative can be computed by:

$$\mathrm{d}\mathbf{x} = (\mathbf{x}_t - \hat{\mathbf{x}}_0)/\sigma_t. \quad (2)$$

We can then obtain the estimated denoised sample $\mathbf{x}_{t-1}$ at the previous time step by:

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \mathrm{d}\mathbf{x}(\sigma_{t-1} - \sigma_t). \quad (3)$$

The above denoising process can be abstracted into two steps: 1) Predicting the clean video given the current noisy latent: $\text{Predict}(\mathbf{x}_t) = \hat{\mathbf{x}}_0$ as defined in Eq. (1); 2) Denoising the current latent to get the previous-time-step latent: $\text{Denoise}(\mathbf{x}_t, \hat{\mathbf{x}}_0) = \mathbf{x}_{t-1}$ as defined in Eqs. (2) and (3). By repeating the two steps, SVD progressively denoises the latent and finally produces a clean video $\mathbf{x}_0$.

3.2. Rendering Artifact-prone Videos

Given multiple training views and an extrapolative novel view lying far from the training views, a radiance field can be trained with techniques like 3D Gaussian Splatting [20] and a video can be further rendered that starts from the nearest training view and gradually transitions to the extrapolative novel view. When only a single view or monocular video is available, depth can be estimated by using off-the-shelf image or video depth estimators such as UniDepth [34] or DepthCrafter [17]. With the estimated depth, the image or monocular video can be projected into a point cloud for rendering a video starting from the initial view to the extrapolative novel view.

The initial video frames usually exhibit a clean and accurate appearance since the rendered video starts from one observed training view. However, significant artifacts and unnatural looking appear as the view of the rendered video

Algorithm 1: Video refinement with guidance annealing and resampling annealing.

Input: artifact-prone video $\tilde{\mathbf{x}}$, opacity mask $\mathbf{m}$

1 $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$
2 **for** $t = T, \ldots, 1$ **do**
3 **if** $t > T - T^{\text{guide}}$ **then**
4 **for** $r = 1, \ldots, R$ **do**
5 $\hat{\mathbf{x}}_0 = \text{Predict}(\mathbf{x}_t)$
6 **if** $r \leq R^{\text{guide}}$ **then**
7 $\hat{\mathbf{x}}_0^{\text{dir}} = \tilde{\mathbf{x}} \odot \mathbf{m} + \hat{\mathbf{x}}_0 \odot (1 - \mathbf{m})$
8 **else**
9 $\hat{\mathbf{x}}_0^{\text{dir}} = \hat{\mathbf{x}}_0$
10 $\mathbf{x}_{t-1} = \text{Denoise}(\mathbf{x}_t, \hat{\mathbf{x}}_0^{\text{dir}})$
11 **if** $r < R$ **then**
12 $\mathbf{x}_t \sim \mathcal{N}(\hat{\mathbf{x}}_0^{\text{dir}}, \sigma_t)$
13 **else**
14 $\hat{\mathbf{x}}_0 = \text{Predict}(\mathbf{x}_t)$
15 $\mathbf{x}_{t-1} = \text{Denoise}(\mathbf{x}_t, \hat{\mathbf{x}}_0)$
16 **return** $\mathbf{x}_0$

frames extends beyond the range of the training views. Nevertheless, the rendered videos still retain valuable information about the scene's geometry and appearance. Given that SVD is trained with large-scale natural videos, we exploit the distribution of natural videos in SVD to inpaint and refine the rendered artifact-prone videos.

3.3. Guidance with Input Videos

Given the rendered artifact-prone video $\tilde{\mathbf{x}}$, our goal is to refine it for a more natural appearance, reducing artifacts while preserving the original content. Since the first frame of $\tilde{\mathbf{x}}$ contains minimal artifacts, it can effectively serve as the image condition for SVD. Beyond the image condition, we also need to condition SVD on the remainder of the video to ensure that the output video retains the original content, including camera movement, scene dynamics, and geometry. We can interpret Eq. (2) as denoising the noisy latent at each time step towards the direction of the predicted clean video $\hat{\mathbf{x}}_0$. To guide the denoising process towards $\tilde{\mathbf{x}}$, we can replace the $\hat{\mathbf{x}}_0$ in Eq. (2) with $\tilde{\mathbf{x}}$. However, since $\tilde{\mathbf{x}}$ may contain regions of the scene that are not fully captured, we also need to leverage SVD for multi-view consistent video inpainting. This can be achieved by allowing SVD to denoise the unseen parts without the guidance from $\tilde{\mathbf{x}}$. Given the opacity mask $\mathbf{m}$ indicating the unseen parts, we can obtain the denoising direction as:

$$\hat{\mathbf{x}}_0^{\text{dir}} = \tilde{\mathbf{x}} \odot \mathbf{m} + \hat{\mathbf{x}}_0 \odot (1 - \mathbf{m}), \tag{4}$$

where the seen parts are used to guide the denoising process, and the unseen parts are inpainted by SVD. Then we can replace the denoising direction in the original denoising step to achieve guided denoising:

$$\mathbf{x}_{t-1} = \text{Denoise}(\mathbf{x}_t, \hat{\mathbf{x}}_0^{\text{dir}}). \tag{5}$$

Methods	SSIM ↑	PSNR ↑	LPIPS ↓
3DGS	0.416	14.46	0.429
DRGS	0.406	14.68	0.457
ViewExtrapolator (video)	0.427	14.83	0.379
ViewExtrapolator (3DGS)	0.460	15.46	0.378
ViewExtrapolator w/o GA	0.442	15.14	0.448
ViewExtrapolator w/o RA	0.456	15.33	0.382

Table 1. **Quantitative comparisons and ablation studies.** The first four rows present the comparison results, while the last two rows show the ablation studies. ViewExtrapolator w/o GA denotes results without guidance annealing, and ViewExtrapolator w/o RA denotes results without resampling annealing.

3.4. Video Refinement

Guidance annealing. While the denoising process in Eq. (5) is guided by the artifact-prone video $\tilde{\mathbf{x}}$, it alone cannot remove the artifacts within $\tilde{\mathbf{x}}$ which predominantly exist in the finer details of the video. Since the diffusion models gradually add details during the denoising process, we guide the denoising process in Eq. (5) during the first T^{guide} denoising steps only, as indicated in line 3 of Algorithm 1. During the rest unguided steps of the denoising process, SVD remains conditioned on the first frame of $\tilde{\mathbf{x}}$ and continues denoising the latent produced after T^{guide} guided steps. This approach allows SVD to generate natural video details based on the clean first frame while retaining the coarse structure from the previously denoised latent, thus reducing the artifacts contained in $\tilde{\mathbf{x}}$ and generating more natural and consistent details.

Resampling annealing. However, artifacts in the latent accumulate during the first T^{guide} denoising steps (as each guided step with $\tilde{\mathbf{x}}$ introduces artifacts), which could become too dominant for SVD to refine in the subsequent unguided denoising steps. Therefore, it is necessary for SVD to refine the denoised latent throughout the initial T^{guide} denoising steps as well. Drawing inspiration from the resampling technique [29] that reduces artifacts by repeating a denoising step multiple times, we incorporate R resampling steps at each of the T^{guide} denoising steps, as indicated in line 4 of Algorithm 1. Specifically, during each guided denoising step t, after obtaining the denoised latent $\mathbf{x}_{t-1}$ from the previous time steps with Eq. (5), we diffuse $\mathbf{x}_{t-1}$ back to $\mathbf{x}_t$ as $\mathbf{x}_t \sim \mathcal{N}(\hat{\mathbf{x}}_0^{\text{dir}}, \sigma_t)$. This is followed by another round of denoising over $\mathbf{x}_t$ as defined in Eq. (5). However, the resampling technique in [29] is originally designed for

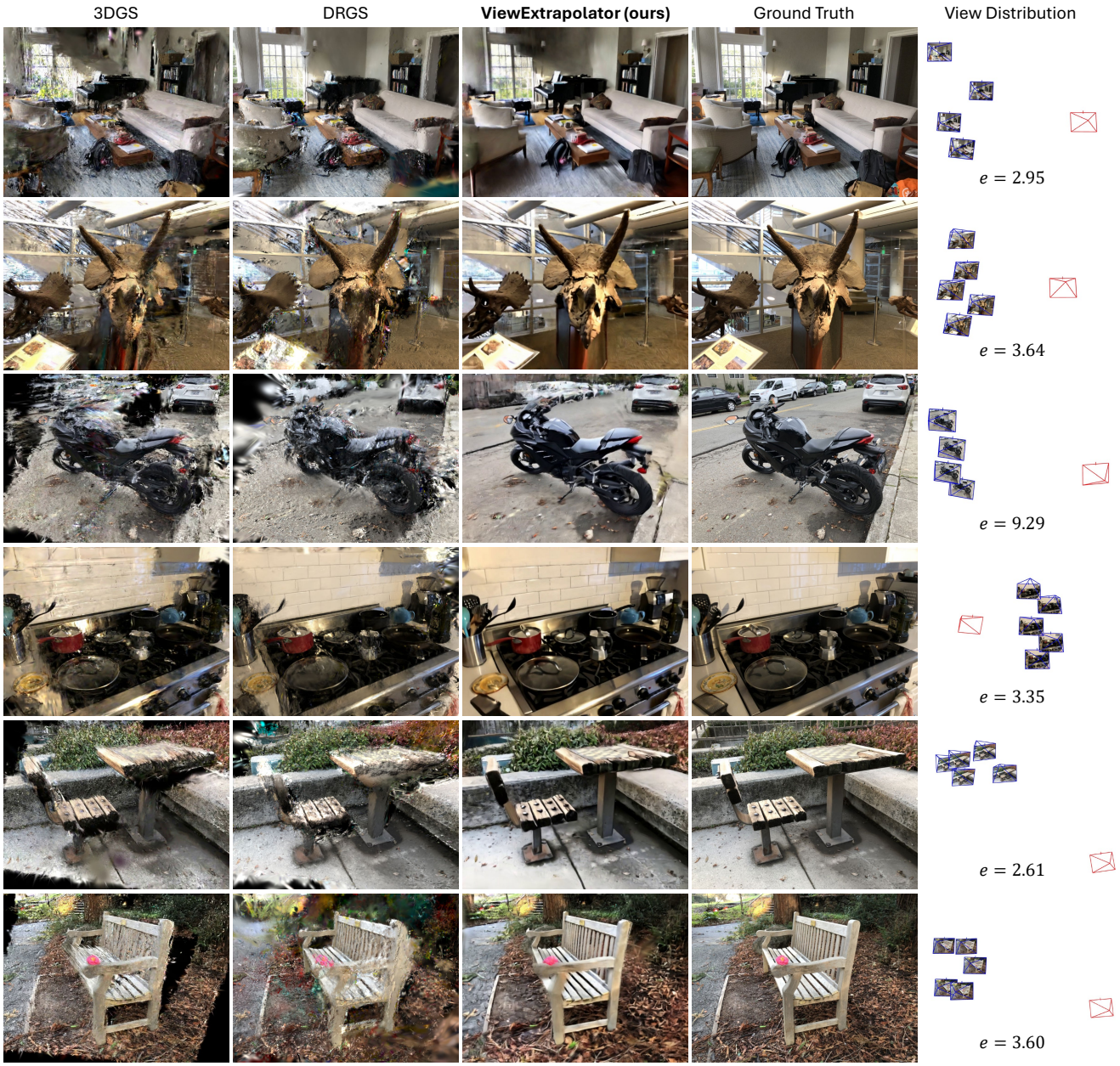

Figure 4. **Qualitative comparisons.** We compare ViewExtrapolator with 3DGS and DRGS on novel view extrapolation. ViewExtrapolator demonstrates superior generation quality with much fewer artifacts. The last column shows the distribution of training and test views as well as the corresponding extrapolation degree e. Zoom in for details.

inpainting tasks where the goal is to preserve the visible regions unaltered, whereas we need to refine artifacts in the visible regions. Since SVD can denoise the latent towards the direction of a natural video that contains few artifacts, we apply the guidance in Eq. (5) only for the first R^{guide} resampling steps in each denoising step, allowing SVD to denoise without the guidance of $\tilde{\mathbf{x}}$ in the remaining resampling steps, as indicated in line 6 of Algorithm 1. During these unguided resampling steps, SVD denoises the latent towards a more natural video, effectively reducing the artifacts introduced in the guided steps.

The above guidance annealing and resampling annealing can be combined and formulated as:

$$\hat{\mathbf{x}}_0^{\text{dir}} = \begin{cases} \hat{\mathbf{x}}_0, & \text{if } t \leq T - T^{\text{guide}} \text{ and } r > R^{\text{guide}} \\ \tilde{\mathbf{x}} \odot \mathbf{m} + \hat{\mathbf{x}}_0 \odot (1 - \mathbf{m}), & \text{else} \end{cases}, \quad (6)$$

where $t \in [1, T]$ is the denoising time step and $r \in [1, R]$

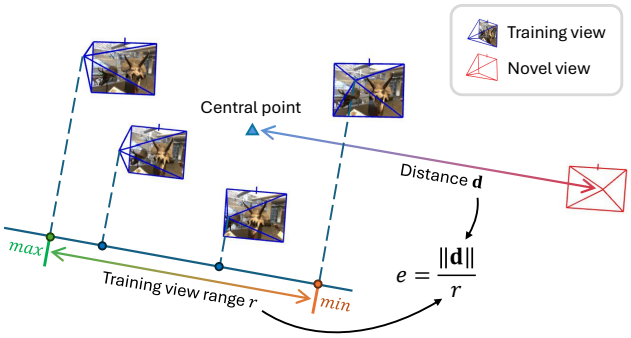

Figure 5. The definition of **extrapolation degree** e by the ratio between $\mathbf{d}$ and r ($\mathbf{d}$ stands for the distance between the novel view and the central point of training views, and r stands for the training view range as the maximum extent of the training views along the direction of $\mathbf{d}$). A higher e means that the novel view is farther away from the training views.

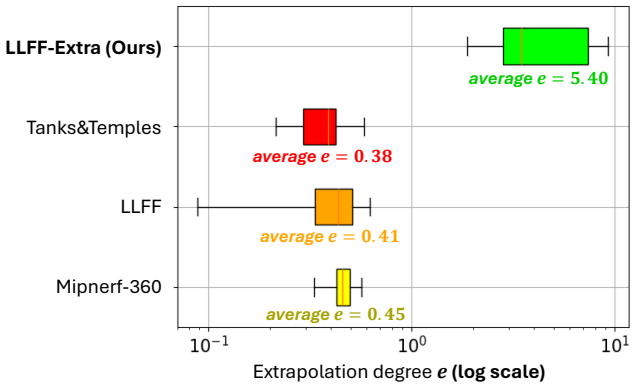

Figure 6. **Distributions of extrapolation degree** e across existing benchmarks and our proposed LLFF-Extra. Unlike LLFF-Extra, all existing benchmarks exhibit a small e, indicating that they predominantly focus on the evaluation of novel view interpolation instead of extrapolation.

is the resampling step. With the guidance from the artifact-prone video and the video refinement with guidance annealing and resampling annealing, we derive the complete denoising algorithm as illustrated in Algorithm 1.

4. Experiments

We conduct extensive experiments to evaluate the proposed ViewExtrapolator on novel view extrapolation. For 3D renderings from radiance fields, we describe the settings of the evaluation dataset in detail (Sec. 4.1) and benchmark ViewExtrapolator with existing methods both qualitatively and quantitatively (Sec. 4.2). For 3D renderings with point clouds, since novel view synthesis from a single view and monocular video is inherently under-constrained, we focus on qualitative evaluations only for highlighting the broad applicability of our method (Sec. 4.3). In addition, we conduct ablation studies to validate the necessity and effectiveness of our key design choices (Sec. 4.4). The implementation details are provided in the appendix.

4.1. Dataset

Effective evaluation of novel view extrapolation requires a dataset where the test views lie significantly beyond the training views for each scene. To create such a dataset, it is crucial to define a metric that can quantify and measure the distance of a novel view from a set of training views. This metric should increase as the novel view moves further away from the training views. In addition, it should be invariant to the scene scale, as camera poses of real-world data are often scaled arbitrarily [38]. To this end, we formulate an intuitive metric called extrapolation degree e as illustrated in Fig. 5. Given a set of training views $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$ and a test novel view $\mathbf{q}$ with similar viewing directions, the distance $\mathbf{d}$ from $\mathbf{q}$ to the centroid of $\mathbf{P}$ can be computed by: $\mathbf{d} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{p}_i - \mathbf{q}$. Another parameter r measuring the range of $\mathbf{P}$ can be derived by the maximum extent of $\mathbf{P}$ along the direction of $\mathbf{d}$ as follows:

$$r = \max_i(\mathbf{p}_i \cdot \frac{\mathbf{d}}{\|\mathbf{d}\|}) - \min_i(\mathbf{p}_i \cdot \frac{\mathbf{d}}{\|\mathbf{d}\|}). \tag{7}$$

The extrapolation degree e can thus be defined by:

$$e = \frac{\|\mathbf{d}\|}{r}. \tag{8}$$

The defined extrapolation degree e thus increases proportionally with $\|\mathbf{d}\|$ when the novel view moves further away from the training views and inversely with r when the training views have more extensive coverage of the scene. It also ensures that the novel view lies outside the convex hull of the training views when $e > 1$. Thus, a novel view with $e > 1$ will likely be in the novel view extrapolation setting.

Most existing benchmarks such as LLFF [31] and Mipnerf-360 [2] are not suitable for evaluating novel view extrapolation as they take an interpolation setting (with small e) by default as illustrated in Fig. 6. We construct LLFF-Extra, a new benchmark that has large e and can be straightly employed to evaluate novel view extrapolation. Specifically, we use 12 scenes from LLFF and select the training views and test novel views with $e = 5.4$ on average, leading to the first benchmark that can be adopted in the future study of novel view extrapolation.

4.2. Benchmarking

We benchmark ViewExtrapolator with the original 3D Gaussian Splatting (3DGS) [20] and its depth-regularized variant DRGS [8] which incorporates depth [3] as a geometric prior to enhance the reconstruction quality. By using 3DGS renderings as the artifact-prone videos, we employ

(a) 3D Gaussian Splatting

(b) Instant NGP

(c) Point cloud from single view

(d) Point cloud from monocular video

Figure 7. **Results from different rendering methods.** Our method can refine view sequences rendered from **(a)** 3D Gaussian Splatting, **(b)** Instant-NGP, and point cloud from **(c)** a single view or **(d)** monocular video. (The top row in each section is the rendered artifact-prone video and the bottom row is the refined video.)

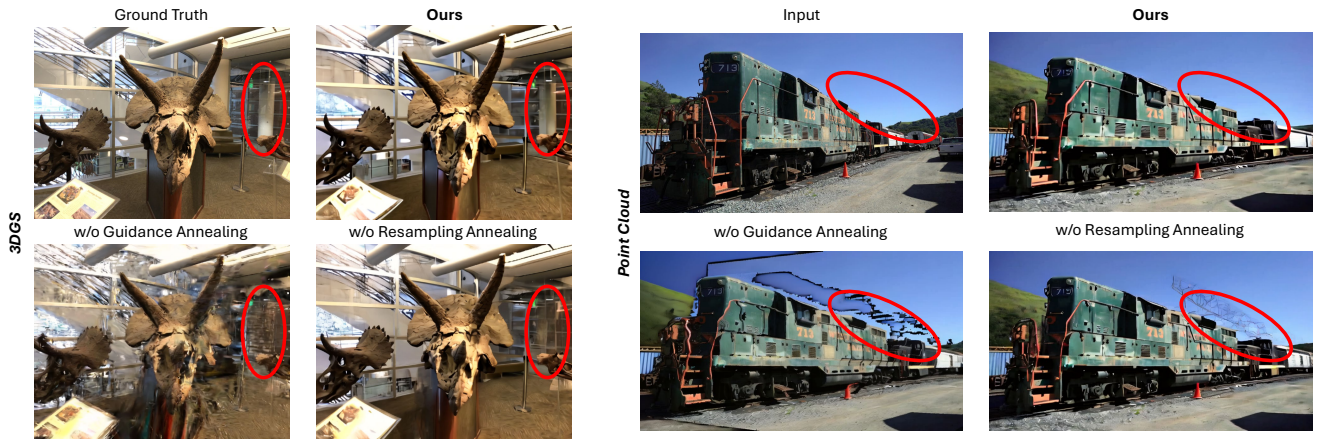

Figure 8. **Ablation studies.** We show the ablation results for 3DGS and point cloud renderings. As point clouds are used for single-image novel view extrapolation without ground truth, we show the input image for reference instead. As highlighted in the red circles, both guidance annealing and resampling annealing are essential for artifact refinement. Please zoom in for details.

the refined video frames (Ours (video) in Tab. 1) to tune the pre-trained 3DGS model and evaluate renderings from the tuned 3DGS model (Ours (3DGS) in Tab. 1) for fair comparison. The quantitative evaluations involve standard novel view synthesis metrics including SSIM, PSNR, and LPIPS [56]. We would highlight that LPIPS is more suitable for evaluating novel view extrapolation which is more toward a generative instead of regressive task with many unseen parts to generate in extrapolative views.

ViewExtrapolator surpasses 3DGS and DRGS both qualitatively and quantitively, achieving superior visual reconstruction with much fewer artifacts as illustrated in Fig. 4 and Tab. 1. One key observation is that 3DGS renderings degrade severely under the novel view extrapolation setting. Additionally, the incorporation of depth priors in DRGS does not lead to much improvement. Both experiments underscore that the core challenge in novel view extrapolation lies with the lack of observations in extrapolated views and direct incorporation of geometry priors alone will not solve the problem. As a comparison, ViewExtrapolator achieves substantial improvement in perceptual quality (LPIPS), demonstrating the effectiveness of novel view refinement with generative priors from SVD.

4.3. Broad Applicability

The proposed ViewExtrapolator is versatile and can generalize to various 3D rendering approaches that often come with different types of artifacts in novel view extrapolation. We verify this feature over renderings by radiance fields and point clouds. For radiance fields, we test ViewExtrapolator over Instant-NGP [33]. Unlike 3DGS artifacts with noisy clusters of 3D Gaussians, Instant-NGP often produces blurry and fine-grained artifacts. ViewExtrapolator corrects both types of artifacts effectively as illustrated in Fig. 7 (a, b). For point clouds, we evaluate ViewExtrapolator over point-cloud renderings when only a single view or monocular video is available. As Fig. 7 (c,d) shows, ViewExtrapolator removes the unique point artifacts effectively. The above studies demonstrate the superior generalization and flexibility of ViewExtrapolator, highlighting its broad applicability across various scenarios with little tuning.

4.4. Ablation Studies

We conduct ablation studies to examine how the proposed guidance annealing and resampling annealing contribute to novel view extrapolation. In the studies, we apply guidance at every diffusion time step and resampling step, respectively, for verifying guidance annealing and resampling annealing. As Fig. 8 and Tab. 1 show, only partial artifacts are refined without resampling annealing while most artifacts remain intact without guidance annealing. This verifies the crucial role of artifact refinement with guidance annealing and resampling annealing.

5. Conclusion

We present ViewExtrapolator, a novel and training-free approach for novel view extrapolation. While current radiance field methods struggle to synthesize novel views that lie far beyond the range of the training views, ViewExtrapolator is able to render realistic views by leveraging the generative priors of SVD. We refine the artifact-prone views rendered by radiance fields by guiding SVD to preserve the scene content and eliminate the artifacts at the same time. ViewExtrapolator demonstrates superior novel view extrapolation quality compared to current methods and can also be applied to point cloud renderings when only a single view or monocular video is available.

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. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2, 6

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 6

[4] 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. 2, 3

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2

[6] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 3

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 2

[8] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2024. 6

[9] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 2

[10] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[11] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 3

[12] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 3

[13] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[15] Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5043–5052, 2024. 3

[16] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3

[17] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 3

[18] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 3

[19] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 3

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2, 3, 6

[21] Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6775–6785, 2024. 3

[22] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3

[23] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8338–8348, 2023. 2

[24] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt,

Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023. 2

[25] Kunhao Liu, Fangneng Zhan, Muyu Xu, Christian Theobalt, Ling Shao, and Shijian Lu. Stylegaussian: Instant 3d style transfer with gaussian splatting. *arXiv preprint arXiv:2403.07807*, 2024. 2

[26] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3

[27] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[28] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 2

[29] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2, 4

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

[31] 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)*, 38(4):1–14, 2019. 6

[32] 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

[33] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 2, 8

[34] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 3

[35] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

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

[37] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 3

[38] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[39] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[40] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[41] Meng-Li Shih, Wei-Chiu Ma, Lorenzo Boyice, Aleksander Holynski, Forrester Cole, Brian Curless, and Janne Kontkanen. Extranerf: Visibility-aware view extrapolation of neural radiance fields with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20385–20395, 2024. 2, 3

[42] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 2

[43] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[44] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 3

[45] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3

[46] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 3

[47] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[48] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[49] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 18120–18130, 2023. 3

[50] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[51] 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. 3

[52] Jamie Wynn and Daniyar Turmukhambetov. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4180–4189, 2023. 3

[53] 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

[54] Jian Zhang, Yuanqing Zhang, Huan Fu, Xiaowei Zhou, Bowen Cai, Jinchi Huang, Rongfei Jia, Binqiang Zhao, and Xing Tang. Ray priors through reprojection: Improving neural radiance fields for novel view extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18376–18386, 2022. 2

[55] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[56] 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. 8

Novel View Extrapolation with Video Diffusion Priors

Appendix

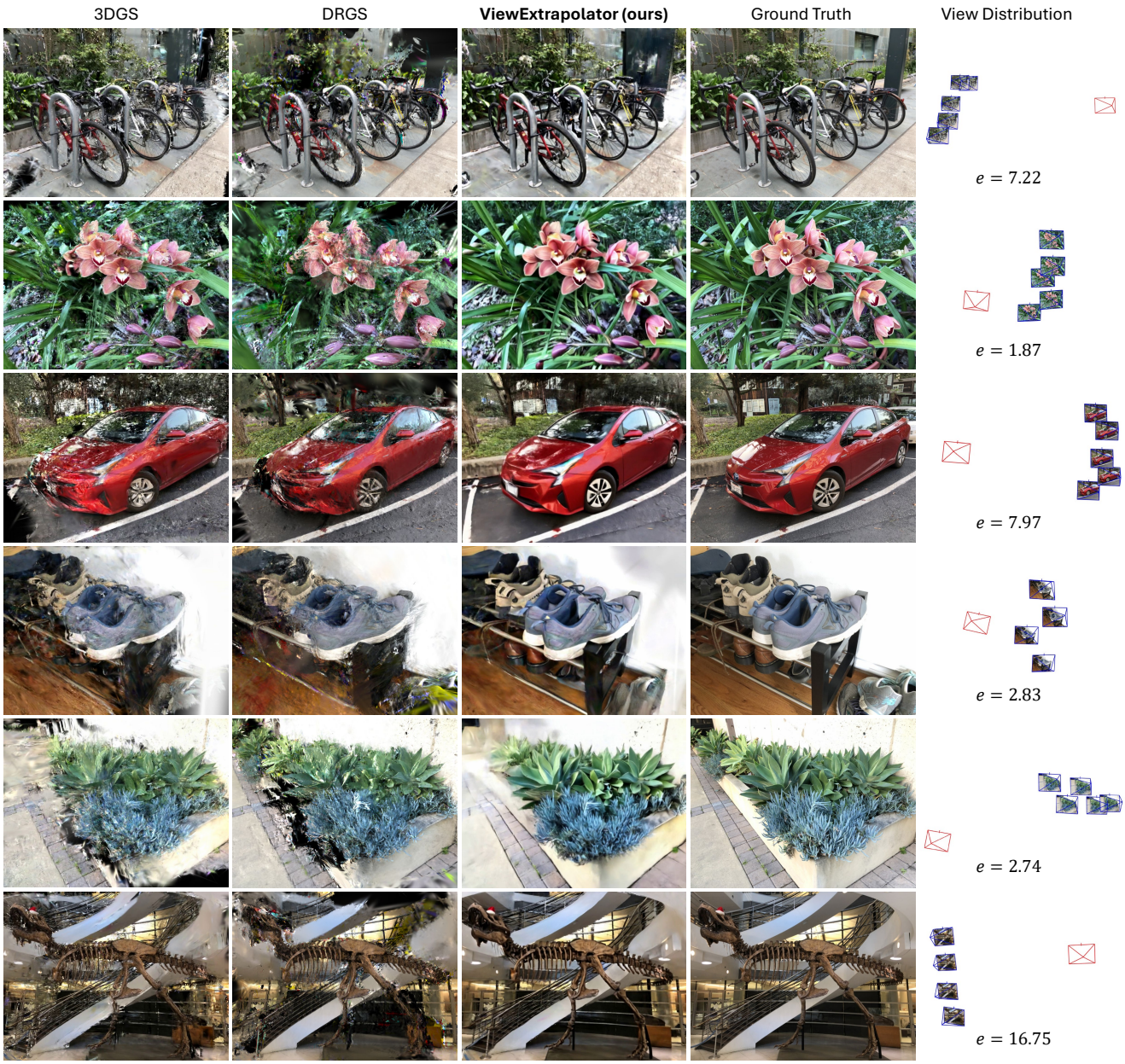

Figure 9. **Additional comparisons.** We compare ViewExtrapolator with 3DGS and DRGS on novel view extrapolation. ViewExtrapolator demonstrates superior generation quality with much fewer artifacts. The last column shows the distribution of training and test views as well as the corresponding extrapolation degree e. Zoom in for details.

A. Additional Results

We show additional qualitative comparisons in Fig. 9. Please visit the project page for video results: https://kunhao-liu.github.io/ViewExtrapolator/.

B. Implementation Details

Hyperparameters. We base our approach on the `xt-1-1` version of the SVD model, which generates 25-frame 6-fps videos at a resolution of 576×1024. For

(a) Novel view from extreme angle

(b) Video with rapid motion

Figure 10. **Limitations and failure cases.** The generation quality would degrade when handling **(a)** novel views at extreme angles or **(b)** dynamic videos with rapid motion. (The top row in each section is the rendered artifact-prone video and the bottom row is the refined video.)

all experiments, we set $T = 25$, $R = 3$, and $R^{\text{guide}} = 1$, with $T^{\text{guide}} = 15$ for static scenes and $T^{\text{guide}} = 16$ for dynamic scenes. We set `noise_aug_strength` $= 0$ to preserve the original scene content and set other parameters as default. Our experiments were conducted on an NVIDIA RTX A5000 GPU with 24G memory, with each video refinement taking 3 minutes and 20 seconds.

Details on 3DGS Refinement. For the evaluations of novel view extrapolation, we employ the refined video frames (Ours (video) in Tab. 1) to tune the pre-trained 3DGS model and evaluate renderings from the tuned 3DGS model (Ours (3DGS) in Tab. 1). Given the refined video frames, we use them as well as the original training views to refine the 3DGS model using the standard L1, SSIM loss, and default densification strategy. In order to let the refined video frames regularize the geometry of 3DGS instead of being fitted as the view-dependent color, we incrementally increase the order of the spherical harmonics during refinement, starting from 0. In addition, to make the refined 3DGS more faithful to the original training views, we gradually decrease the frequency of training iterations that use the refined video frames throughout the training process. The refinement process requires one-third of the iterations used in the original 3DGS training.

C. Limitations

Although ViewExtrapolator offers advantages in novel view extrapolation, it has several limitations. First, as an inference-stage approach, the quality ceiling of our method is bound by the original SVD model, meaning it also inherits certain drawbacks, such as lower resolution and color shifts. We believe incorporating more advanced video diffusion models could help enhance the overall quality. Second, our method encounters challenges when handling dynamic videos with rapid motion or extreme views where the novel views have very little overlap with the observed scene. We show the limitations and failure cases in Fig. 10.