Enhancing Novel View Synthesis via Geometry Grounded Set Diffusion

Farhad G. Zanjani **Hong Cai** **Amirhossein Habibian**

Qualcomm AI Research*

{fzanjani, hongcai, ahabibia}@qti.qualcomm.com

Abstract

We present *SetDiff*, a geometry-grounded multi-view diffusion framework that enhances novel-view renderings produced by 3D Gaussian Splatting. Our method integrates explicit 3D priors, pixel-aligned coordinate maps and pose-aware Plücker ray embeddings, into a set-based diffusion model capable of jointly processing variable numbers of reference and target views. This formulation enables robust occlusion handling, reduces hallucinations under low-signal conditions, and improves photometric fidelity in visual content restoration. A unified set mixer performs global token-level attention across all input views, supporting scalable multi-camera enhancement while maintaining computational efficiency through latent-space supervision and selective decoding. Extensive experiments on EUVS, Para-Lane, nuScenes, and DL3DV demonstrate significant gains in perceptual fidelity, structural similarity, and robustness under severe extrapolation. SetDiff establishes a state-of-the-art diffusion-based solution for realistic and reliable novel-view synthesis in autonomous driving scenarios. https://qualcomm-ai-research.github.io/setdiff

1 Introduction

The recent paradigm shift toward end-to-end intelligent systems based on vision–language–action models [14, 18, 30] has created a growing demand for photo-realistic simulators, or world models, that can support closed-loop training and evaluation. While fully diffusion-based simulators have made significant progress [1, 31, 8], they often struggle to maintain the temporal consistency and cross-view fidelity required for realistic autonomous driving simulation. An alternative and increasingly prominent direction is to leverage 3D Gaussian Splatting (GS) [13] for sensor simulation [41, 27, 44, 21]. GS-based methods offer compelling advantages in temporal coherence, geometric fidelity, and computational efficiency. However, they remain vulnerable to distribution shifts in camera viewpoint: when the agent deviates from trajectories observed during training—a frequent occurrence in closed-loop settings, as shown in Figure 2—rendering quality can degrade sharply. This limitation makes robust novel view synthesis (NVS) a key bottleneck for deploying GS-based simulators in robotics and autonomous driving. In this work, we close this gap in NVS quality by improving the robustness of GS-based rendering under challenging *autonomous driving* scenarios, particularly when the ego agent deviates from training trajectories.

Prior efforts to improve NVS can be broadly grouped into three categories. *i)* Geometry-regularized reconstruction methods [16, 7, 5] enforce depth, normal, or structural constraints to promote geometric consistency. While effective to some extent, these methods remain sensitive to noisy supervision, sparse viewpoints, and wide-baseline extrapolations, often leading to brittle or unstable reconstructions. *ii)* Optimization methods that inject generative priors, such as RegNeRF [26] and VEGS [11], can hallucinate plausible content for unseen regions, but often rely on single-view supervision or

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

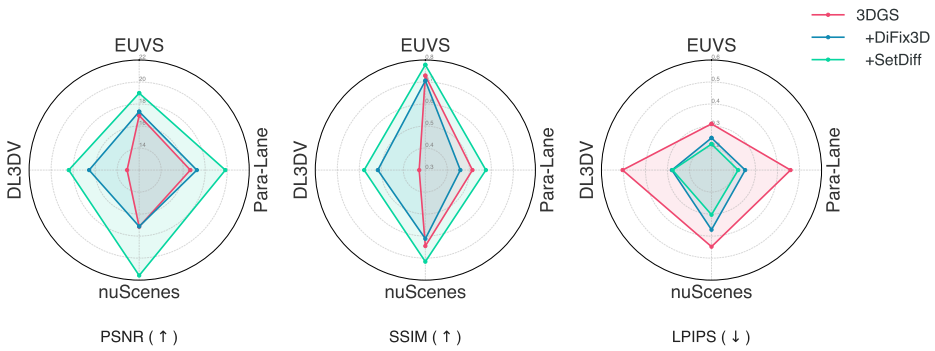

Figure 1: Our proposed SetDiff consistently improves 3DGS [13] performance for NVS, on four datasets, while outperforming DIFIX [35] as the SOTA diffusion-based enhancer.

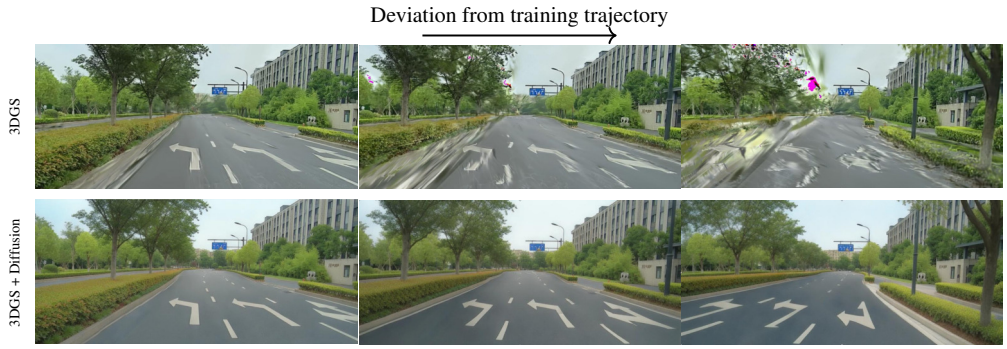

Figure 2: **3DGS novel-view extrapolation** exhibits increasing artifacts with trajectory deviation—less than one lane (left), one lane (middle), and two lanes (right). Our diffusion-based enhancement suppresses artifacts in the bottom row and preserves quality under extreme extrapolation.

static priors, which limits multi-view consistency and temporal coherence. *iii)* Most relevant to our setting are diffusion-based enhancement methods [20, 35]. In particular, DiFix3D [35] formulates NVS as an image-to-image diffusion enhancement problem and trains its model on carefully curated pairs of noisy renderings and clean ground-truth images. While these methods can produce visually appealing outputs, they condition primarily on RGB renderings and rely heavily on the diffusion prior to hallucinate corrupted regions. As a result, they lack explicit geometric guidance and struggle to effectively incorporate reference image evidence. When input signal-to-noise ratios are low, the enhanced results may appear perceptually compelling but often deviate from the true underlying scene in terms of photometric fidelity.

We address these limitations with *SetDiff*, a geometry-grounded diffusion enhancer designed to process an arbitrary-length, arbitrarily ordered *set* of images rendered from diverse camera poses and time steps. Our model employs an orderless *set-mixer* attention mechanism that enables the diffusion process to attend to and aggregate complementary information across viewpoints, thereby maximizing enhancement quality. To ensure efficiency while capturing cross-frame and cross-view dependencies, we construct the diffusion denoiser using a 2D U-Net architecture distilled into a single-step inference model. This design yields a highly efficient enhancer without sacrificing the ability to leverage rich multi-view geometric context.

Crucially, rather than relying solely on RGB renderings, we introduce a *geometry-grounded* conditioning strategy that exposes the denoiser to explicit 3D scene geometry extracted from the 3DGS representation. Specifically, we condition the diffusion model on rendered coordinate maps that provide pixel-aligned 3D correspondences, along with pose-aware embeddings, such as Plücker ray encodings. These geometric cues enable the denoiser to reason about where geometry is consistent, uncertain, or conflicting across views, allowing it to suppress floaters, respect occlusions, and recover high-frequency details without compromising structural plausibility. We experimentally demonstrate

that such geometry grounding substantially improves content restoration and overall novel view synthesis quality. Our contributions are summarized as follows:

- We introduce a geometry-grounded conditioning strategy that renders 3D cues from Gaussian Splatting representations and injects them into the diffusion denoiser, improving content fidelity and reducing hallucinations under degraded RGB renderings.
- We propose a variable-length diffusion architecture that jointly processes a set of reference and target views, enabling effective content propagation across images for stronger multi-view visual aggregation. Owing to its scalable design and training strategy, our model supports near-online enhancement of high-resolution images.
- We establish new state-of-the-art results in novel view synthesis on three challenging autonomous driving datasets—EUVS, Para-Lane, and nuScenes—as well as on the general-scene benchmark DL3DV as summarized in Figure 1.

2 Related Work

2.1 Neural Rendering and Gaussian Splatting for Autonomous driving: Neural scene representations—such as Neural Radiance Fields (NeRF) [23] and in particular 3D Gaussian Splatting (3DGS) [13]—have recently been extended to autonomous driving scenarios [6, 39, 42, 46, 37, 43]. While these approaches substantially improve photorealism and support reconstruction of large-scale driving scenes, they remain fundamentally limited by the shortcomings of their underlying representations in novel view synthesis. Their performance degrades markedly under sparse or uneven sensor coverage and during wide-baseline viewpoint extrapolation, often resulting in artifacts such as floaters, geometry inconsistencies, and texture bleeding. Such limitations hinder their viability for direct use in generalizable closed-loop simulation pipelines, where stable view synthesis, robustness to sparse observations, and camera-pose extrapolation are essential.

2.2 Generative Priors in 3D Optimization: To alleviate reconstruction artifacts, several approaches incorporate generative models as priors during optimization. RegNeRF [26] and DietNeRF [12] introduce semantic and photometric regularization to guide appearance learning. ReconFusion [36] and VEGS [11] leverage pretrained generative models to hallucinate unseen content and improve novel view synthesis. In particular, VEGS distills diffusion-based priors into 3DGS to support urban-scale view extrapolation. While these techniques improve reconstruction by introducing generated unseen views as pseudo-supervision during optimization, they often struggle in severely degraded settings and typically do not ensure strong content fidelity and multi-view coherence.

2.3 Diffusion-Based Enhancement: Another line of work treats novel-view enhancement as image-level restoration with diffusion models. DiFix3D [35] finetunes a text-to-image latent diffusion model to enhance rendered views by treating them as noisy inputs; however, its RGB-only conditioning can lead to over-hallucination and reduced scene fidelity. DiFix++ [28] further improves restoration quality with refined objectives. 3DGS-Enhancer [20] enhances intermediate rendered frames between two high-quality reference views using a video diffusion prior for temporal coherence. However, its single-camera formulation and sequential generation pipeline limit scalability to multi-view and multi-camera simulation settings.

Inspired by prior work, we introduce a set-based diffusion enhancer that jointly restores multi-view renderings using a distilled diffusion prior. In contrast to existing approaches, our framework naturally handles variable numbers of cameras and time steps, and incorporates explicit geometric conditioning through occlusion-aware cross-view correspondences and known camera poses. This design enables robust enhancement even under severe degradation by transferring missing information from diverse and relevant reference views into novel target views, ultimately improving both perceptual quality and structural fidelity.

3 Geometry Grounded Set Diffusion

Given a 3DGS model trained on a set of images from seen camera trajectories, called the *reference set*, our goal is to enhance the visual quality when rendering the scene from novel camera poses $\{\mathbf{p}_j^t\}_{j=1}^M$. These rendered images form the *target set* $\mathcal{I}^t = \{\hat{\mathbf{I}}_j^t\}_{j=1}^M$. Due to incomplete observations, the target set often suffers from severe rendering artifacts, especially when the novel camera poses deviate

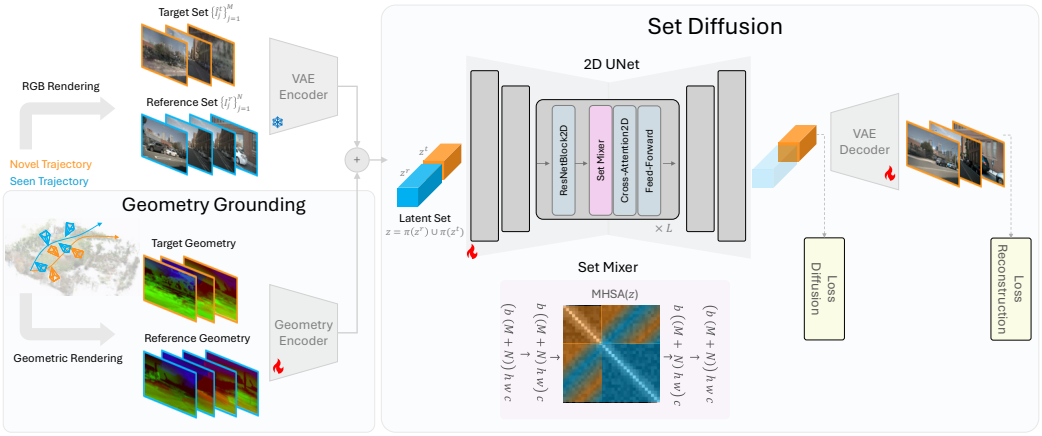

Figure 3: Overview illustration of SetDiff. The rendered novel-view images are enhanced via an image diffusion model, conditioned on reference views, camera poses and 3D geometric priors.

further from the seen trajectories. Following the recent trend [20, 35, 28], we rely on diffusion-based image enhancement to remove novel view rendering artifacts. We propose two major extensions to prior diffusion enhancers: *i)* We ground the diffusion model on rich 3D scene geometries to overcome the lack of explicit 3D modeling in image diffusion models, as described in Section 3.1. *ii)* Instead of processing the target frames and reference frames one by one, we extend the diffusion model to process a variable number of target frames and reference frames jointly as a set, as described in Section 3.2. This allows a more effective flow of information across target and reference frames leading to higher quality enhancements. An overview of our approach is shown in Figure 3.

3.1 Geometric Grounding

Being trained on web-scale images, the existing image diffusion models lack an explicit understanding of 3D structures. The lack of 3D understanding limits their effectiveness in correcting severe rendering artifacts, leading to unrealistic and inconsistent generations. As a remedy, we propose to extract the rich scene geometry information—namely, 3D coordinate maps and camera poses—that is readily available in 3DGS and encode it into the diffusion process.

3.1.1 Extracting Coordinate Maps In addition to standard RGB rendering—where color values associated with 3D Gaussian primitives are projected onto the image plane—we compute a complementary *geometric rendering* that projects the 3D world coordinates of Gaussian centers into the camera view.

Whereas RGB rendering provides appearance information, geometric rendering produces per-pixel 3D coordinate maps (C-maps), which encode the estimated *xyz* position of the visible surface point from the camera's perspective. Although C-maps for target views may contain noise due to extrapolation or imperfect geometry, they serve as valuable structural cues during diffusion-based restoration. Conditioning the diffusion model on C-maps for both reference and target views provides 3D-aware guidance for the multi-view enhancement task. Analogous to positional encodings, C-maps supply pixel-aligned world coordinates that help guide attention across an unordered set of frames, including heterogeneous cameras and temporal frames. Figure 4 shows rendered RGB images and C-maps from 3DGS for two scenes from DL3DV data.

Formally, a *3D coordinate map* $\mathbf{c}(u, v) \in \mathbb{R}^3$ assigns to each pixel (u, v) the accumulated world coordinate of the surface point along the corresponding camera ray. The coordinate is obtained through an opacity-weighted volumetric accumulation:

$$\mathbf{c}(u, v) = \sum_i \alpha_i(u, v) \prod_{j=1}^{i-1} \big(1 - \alpha_j(u, v)\big)\, \mathbf{x}_i, \tag{1}$$

where $(u, v) \in \{1, \ldots, H\} \times \{1, \ldots, W\}$ denotes pixel coordinates for an image of resolution $H \times W$. Here, $\mathbf{x}_i \in \mathbb{R}^3$ represents the 3D world coordinate of sample i along the ray (e.g., a Gaussian center), and $\alpha_i(u, v)$ is its effective opacity contribution. This formulation parallels standard volumetric rendering, where samples contribute to the final estimate based on visibility and transparency.

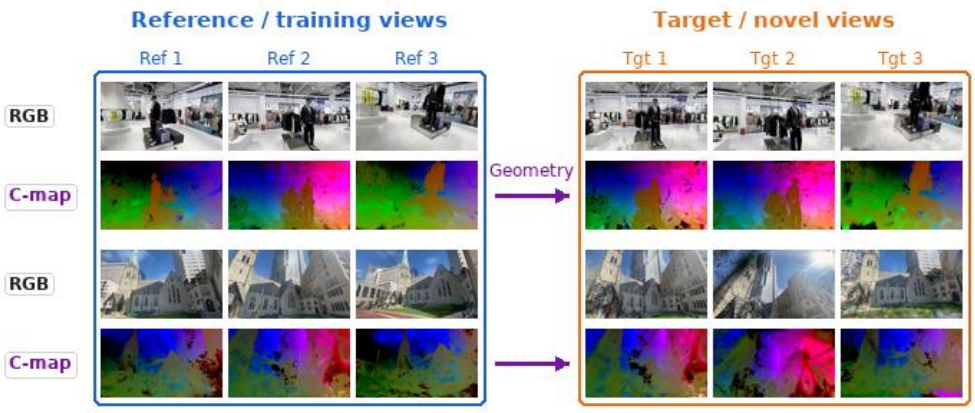

Figure 4: Occlusion-aware correspondence visualization for two DL3DV scenes. The colors in the C-maps encode visibility-aware 3D correspondences between reference and target views. Even when affected by noise from under-reconstructed regions, the maps provide useful geometric guidance to the model.

In 3DGS reconstruction, $\alpha_i(u, v)$ corresponds to the opacity of the i-th Gaussian after projection onto pixel (u, v). An analogous formulation applies to NeRF-style radiance fields when computing coordinate maps.

3.1.2 Extracting Camera Poses Conditioning the enhancement process on camera poses enables the diffusion model to restore target views by effectively transferring and fusing visual information from high-quality reference views. Following prior work [38], we represent camera geometry using *Plücker ray embeddings*, which encode each pixel's viewing ray as a dense 2D feature map. A ray is parameterized by a pair $(\mathbf{o} \times \mathbf{d},\ \mathbf{d}) \in \mathbb{R}^6$, where $\mathbf{o}$ denotes the ray origin in world coordinates, and $\mathbf{d}$ is the normalized ray direction. The cross product $\mathbf{o} \times \mathbf{d}$ captures the ray's moment, forming together with $\mathbf{d}$ a compact and rotation-equivariant representation of 3D lines.

To ensure stable conditioning across heterogeneous camera configurations, we apply a normalization procedure to all camera poses before computing Plücker embeddings. This includes normalizing extrinsic coordinates to a canonical reference. Detailed descriptions of the normalization scheme are provided in the supplementary material.

3.1.3 Geometry Encoding Given the coordinate maps $\mathbf{c}$ and Plücker ray embeddings $\mathbf{p}$, we extract the geometric conditioning tensor $\mathbf{g} \in \mathbb{R}^{H/f \times W/f \times c}$ using a convolutional *geometry encoder* as $\mathbf{g} = \Psi(\mathbf{c}, \mathbf{p})$. The encoder is made of five convolutional blocks with group normalization and SiLU activations. Having the same number of channels and down-sampling factor ($f = 8$) as the VAE encoder ensures that the encoder outputs have matching dimensions, which allows us to combine the photometric and geometric conditioning via element-wise addition. Our ablation studies show that this simple fusion outperforms using cross-attention to consume geometric conditioning.

3.2 Set Diffusion

Following previous work [35, 28], we rely on reference-conditioned image diffusion to enhance the noisy targets. Reference images are defined as a set of clean images retrieved from the seen trajectories, which are closest matches to the novel camera poses.

3.2.1 Reference Set Construction To select the most relevant reference views for the target views, we compute a *view overlap score* between each reference and target camera:

$$\text{score} = \alpha \cdot \frac{v_i \cdot v_j}{\|v_i\| \cdot \|v_j\|} + \beta \cdot \left(1 - \frac{d(c_i, c_j)}{\max_{i,j} d(c_i, c_j)}\right), \tag{2}$$

where v_i and v_j denote the view directions of the ith and j^{th} cameras, respectively, and $d(c_i, c_j)$ is the Euclidean distance between their camera centers, normalized by the maximum pairwise distance in the scene. The parameters α and β control the relative importance of view alignment and spatial proximity. In all experiments, we set $\alpha = 0.8$ and $\beta = 0.2$. It results in a set of N clean reference images per target set, denoted as $\mathcal{I}^r = \{\mathbf{I}^r_j\}_{j=1}^N$.

3.2.2 Latent Construction All images in the reference and target sets, $\mathcal{I}^r$ and $\mathcal{I}^t$, are encoded into the latent space using a frozen image encoder from Stable Diffusion [34]. Geometric conditioning inputs, i.e. C-maps and Plücker ray embeddings, are embedded using a geometry encoder Ψ whose outputs are added to VAE latents. This results in two sets of latents $\mathbf{z}^t \in \mathbb{R}^{M\times H/8\times W/8\times c}$ and $\mathbf{z}^r \in \mathbb{R}^{N\times H/8\times W/8\times c}$ for target and reference sets, respectively.

3.2.3 Set Mixer Unlike previous work [35, 28], which relies on a single reference image and a single target image for diffusion enhancement, we extend the diffusion model to consume a *set* of reference and target images. Using a set of reference images enables the diffusion model to aggregate the information across multiple complementary views rather than solely relying on the closest match, which may miss crucial details needed for an effective enhancement. Moreover, we enhance target frames jointly as a set rather than individually, which improves enhancement quality, as demonstrated experimentally.

Instead of relying on a *vector* representation, which fixes the number and the position of reference and target frames, we rely on a flexible set representation that allows us to use an arbitrary number and an arbitrary order of target and reference images. This flexibility allows a single trained model to handle a variable number of target frames and their matched reference frames and generalize well to arbitrarily ordered reference and target views. For this purpose, we rely on an image-based denoiser with no spatio-temporal convolutions or position encoding in the temporal domain. In particular, we use the 2D UNet from Stable Diffusion, which meets these requirements. Additionally, during the training, we apply a random temporal permutation function π to the target and reference frames, $\pi(\mathbf{z}^t)$ and $\pi(\mathbf{z}^r)$, to increase the model's robustness against temporal reordering.

To effectively aggregate visual information across the reference and target sets, we replace each self-attention layer in the UNet with a *Set Mixer*. It combines all tokens from the target and reference sets, i.e. reshaping into $b \times ((M+N)hw) \times c$ where b denotes batch size, and performs multi-head self-attention (MHSA) over the unified token set as illustrated in Figure 3 . It allows each token from the target or reference images to attend to all other tokens to globally search for the missing details in any reference or target views. As validated experimentally, our set diffusion adds minimal computational overhead to the frame-by-frame processing allowing the batch processing of frames in real time.

3.2.4 Diffusion Training We train our diffusion enhancer following the standard single-step image-to-image diffusion training [29, 35, 28] based on SD-Turbo [34]. Specifically, we fine-tune the denoising UNet f_θ to take the noisy target and clean reference latents, $\mathbf{z}^t$ and $\mathbf{z}^r$, as input and predict the clean target views $\mathbf{y}^t$ as the ground-truth. Following [35], we fix the noise level $\tau = 200$, and adopt the v-prediction objective [33] with the following latent diffusion loss:

$$\mathcal{L}_{\text{Diffusion}} = \mathbb{E}_{\mathbf{z}^r,\mathbf{z}^t,\mathbf{y}^t}\left[\left\|\mathbf{y}^t - f_\theta\big([z^t; z^r], \tau\big)\right\|_2^2\right] \tag{3}$$

We augment the latent diffusion objective with a pixel-space reconstruction loss, $\mathcal{L}_{\text{Reconstruction}}$, enabling joint fine-tuning of the VAE decoder together with the geometry encoder and the denoising U-Net. The reconstruction loss is defined as an equally weighted combination of ℓ_2, SSIM, and LPIPS terms. This formulation jointly enforces pixel-level accuracy, structural consistency, and perceptual fidelity, leading to improved enhancement quality. Given the high memory requirements involved in optimizing the pixel-space reconstruction loss, during training, we only decode a subset of target latents and rely on gradient accumulation and LoRA fine-tuning of the VAE decoder.

4 Experiments

4.1 Experimental Setup

Datasets To evaluate the generalization ability of novel view synthesis (NVS) methods under challenging extrapolation, sparsity, and dynamic scene conditions, we conduct experiments across four diverse datasets: *i)* EUVS [9], a large-scale benchmark designed to evaluate NVS under significant viewpoint changes—crucial for vision-centric autonomous driving. *ii)* Para-Lane [25], a dataset of multi-lane urban driving sequences captured from parallel trajectories, enabling evaluation of NVS under realistic cross-lane displacements. *iii)* nuScenes [4] to evaluate the model's performance under significant scene dynamics, including moving vehicles, pedestrians, and riders. *iv)* DL3DV-10K [19]



Figure 5: **Novel-view images of EUVS (col. 1-4) and Para-Lane (col. 5-6).** The qualitative comparison of 3DGS and enhanced versions by DiFix3D and SetDiff. The SetDiff results show a higher fidelity to the visual content of the scene.

Method	EUVS Novel-View Data Settings								
	Setting 1			Setting 2			Setting 3		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
3DGS [13]	16.37	0.720	0.260	19.53	0.751	0.267	14.99	0.717	0.405
2DGS [10]	16.30	0.710	0.289	18.83	0.720	0.292	11.36	0.545	0.546
GSPro [7]	16.39	0.719	0.245	19.39	0.747	0.225	14.82	0.699	0.388
PGSR [5]	16.32	0.710	0.273	18.38	0.712	0.253	14.25	0.698	0.436
3DGM [17]	16.35	0.725	0.254	18.78	0.746	0.281	14.60	0.723	0.405
Feature 3DGS [45]	16.01	0.724	0.257	19.59	0.786	0.228	14.33	0.639	0.382
Zip-NeRF [3]	14.06	0.692	0.342	17.36	0.672	0.358	14.42	0.656	0.454
Instant-NGP [24]	12.65	0.625	0.594	17.15	0.721	0.517	14.39	0.710	0.659
3DGS [13]	16.37	0.720	0.260	19.53	0.751	0.267	14.99	0.717	0.405
+ VEGS [11]	15.88	0.705	0.306	**23.33**	0.795	0.281	14.25	0.647	0.442
+ DIFIX3D [35]	17.09	0.739	0.224	20.02	0.713	0.187	14.84	0.665	0.328
+ DIFIX3D++ [28]	**17.94**	0.782	0.218	21.68	0.772	0.209	15.91	0.743	0.335
+ **SetDiff**	17.84	**0.796**	**0.192**	22.55	**0.796**	**0.172**	**16.60**	**0.745**	**0.291**

Table 1: **SOTA comparisons on EUVS** across three extrapolation settings. Comparisons of neural reconstruction methods (top) and diffusion-based enhancers (bottom) are reported separately. All diffusion enhancers are applied to renderings from the same 3DGS backbone.

to evaluate the model on diverse bounded and unbounded scenes, beyond the autonomous driving domain. Dataset details and experimental protocols are provided in Appendix A.1.

Evaluation metrics We evaluate the quality of synthesized novel views using PSNR, SSIM, and LPIPS. These metrics quantify complementary aspects of reconstruction quality: PSNR captures pixel-wise fidelity, SSIM measures structural similarity, and LPIPS assesses perceptual consistency. All metrics are computed against ground-truth novel-view images provided by each dataset.

Implementation details During training, we use variable-size target and reference sets with 8–12 frames in total; at inference, the set size is constrained by available GPU memory. Larger reference sets provide richer context, whereas larger target sets enable stronger joint processing and cross-view message passing (as reported in Table 6). C-maps are obtained via global-frame 3DGS rasterization (Equation (1)), clipped to the scene scale, with zero assigned to rays that hit no primitives. No filtering, inpainting, or uncertainty modeling is applied beyond the native 3DGS opacity and Gaussian parameters. We train with classifier-free guidance (CFG), dropping both C-maps and camera poses with probability 15%, and use a CFG scale of 2.0 at inference. All experiments use a resolution of 576×1024. Unless stated otherwise, reference view(s) are provided to diffusion-based enhancers such as DIFIX, 3DGS-Enhancer, and DIFIX++. Further details are provided in the supplement.

4.2 Main Results

We compare SetDiff against state-of-the-art Gaussian Splatting approaches and diffusion-based enhancers for novel view synthesis across four datasets.

EUVS results Table 1 reports the performance of 3D reconstruction baselines and diffusion-based enhancement models evaluated on 3DGS-rendered images. Across all three EUVS novel-view configurations, SetDiff provides the best overall PSNR/LPIPS trade-off among the evaluated diffusion-based enhancers, while not leading every metric or evaluation setting.

The gains are particularly notable in Settings 2 and 3, which involve more challenging extrapolation regimes. Visual comparisons in Figure 5 illustrate how our method restores missing structure and reduces rendering artifacts in extrapolated views, producing higher-quality results compared to competing techniques.

Para-Lane results Table 2 presents results on the Para-Lane dataset. Our method achieves superior performance across both single- and double-lane shifts, reducing LPIPS by approximately 50% relative to raw 3DGS renderings. Compared to other diffusion-based enhancers, SetDiff delivers roughly a 1 dB increase in PSNR while achieving lower LPIPS scores, indicating improved visual realism alongside stronger photometric consistency with the underlying scene. Qualitative examples in Figure 5 demonstrate sharper textures, more coherent geometric details, and robust restoration under cross-lane viewpoint shifts.

Table 2: **SOTA comparisons on Para-Lane** across two extrapolation settings. Comparisons of neural reconstruction methods (top) and diffusion-based enhancers (bottom) are reported separately.

Method	Single-lane transition			Double-lane transition		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
3DGS [13]	17.05	0.524	0.446	16.26	0.505	0.472
2DGS [10]	16.79	0.523	0.469	16.01	0.510	0.494
GSPro [7]	17.01	0.521	0.446	16.29	0.505	0.472
Scaffold-GS [22]	17.59	0.538	0.450	17.09	0.525	0.470
StreetGS [39]	17.50	0.510	0.456	16.16	0.496	0.480
3DGS [13]	17.05	0.524	0.446	16.26	0.505	0.472
+ DIFIX3D [35]	17.62	0.470	0.234	16.87	0.447	0.271
+ DIFIX3D++ [28]	19.25	0.559	0.275	18.55	0.538	0.311
+ SetDiff	**20.20**	**0.586**	**0.208**	**19.47**	**0.564**	**0.233**

Table 3: **SOTA comparisons on nuScenes** reported on extrapolated trajectories.

Method	PSNR	SSIM	LPIPS
OmniRe [6]	17.15	0.645	0.448
+ DIFIX3D [35]	17.12	0.612	0.371
+ DIFIX3D++ [28]	18.74	0.696	0.375
+ SetDiff	**21.58**	**0.716**	**0.303**



Figure 6: **Novel views on nuScenes.** Odd rows show reference and temporally extrapolated OmniRe renderings [6]; even rows show reference-conditioned SetDiff enhancements.

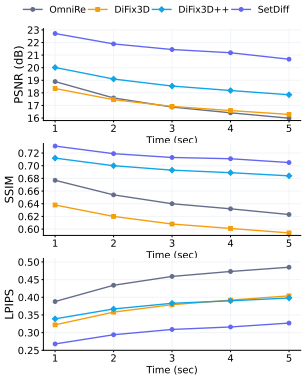

Figure 7: **SOTA comparisons on nuScenes** reported under various trajectory lengths.

nuScenes results Table 3 summarizes novel-view synthesis performance using all six nuScenes cameras over 5-second temporally extrapolated sequences generated by OmniRe [6]. SetDiff substantially improves image quality by correcting severe distortions and reconstructing missing details. As shown in Figure 7, metrics grouped by temporal offset reveal a progressive drop in signal-to-noise ratio (SNR) with longer extrapolation, which is also visually evident in Figure 6. Across all temporal buckets, SetDiff delivers consistently higher fidelity and more realistic reconstructions, effectively restoring information even under long-horizon predictions. Notably, in low-SNR regimes the C-maps become highly noisy and unreliable; however, training SetDiff across varying SNR levels enables the model to exploit C-map conditioning whenever it is reliable.

DL3DV results Table 4 reports performance on DL3DV under three sparse-view configurations. For diffusion-based enhancement models—including DiFix, 3DGS-Enhancer, DiFix++, and the proposed

SetDiff —the 3DGS-rendered novel views are used as input for enhancement. All diffusion-based methods outperform the raw 3DGS outputs, with SetDiff achieving the highest PSNR and lowest LPIPS across all three settings, surpassing DiFix3D and 3DGS-Enhancer. Notably, 3DGS-Enhancer incorporates iterative feedback by reinserting enhanced views into the 3DGS optimization loop, whereas the other methods—including ours—operate purely as post-processors without modifying the underlying geometry. This highlights the effectiveness of SetDiff in improving perceptual quality without requiring retraining, iterative refinement, or geometry updates.

4.3 Ablation Study

To evaluate the contributions of the proposed geometric grounding and set-based diffusion enhancement, we conduct a series of ablation experiments on the DL3DV dataset. Additional ablations are provided in Appendix A.2.1.

Geometric grounding We assess the influence of the proposed geometric conditioning to ground the diffusion enhancer by selectively including camera C-maps, camera poses, or both. As shown in Table 5, using no grounding results in the lowest performance, confirming the importance of explicit geometric priors. Conditioning on either camera poses or C-maps individually yields comparable improvements, indicating that each provides useful geometric grounding. Their combination ("both") achieves the best overall performance, demonstrating that their combination performs best and is complementary.

Method	3 views			6 views			9 views		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
3DGS [13]	10.97	0.248	0.567	13.34	0.332	0.498	14.99	0.403	0.446
Mip-NeRF [2]	10.92	0.191	0.618	11.56	0.199	0.608	12.42	0.218	0.600
RegNeRF [26]	11.46	0.214	0.600	12.69	0.236	0.579	12.33	0.219	0.598
FreeNeRF [40]	10.91	0.211	0.595	12.13	0.230	0.576	12.85	0.241	0.573
DNGaussian [16]	11.10	0.273	0.579	12.67	0.329	0.547	13.44	0.365	0.539
3DGS [13]	10.97	0.248	0.567	13.34	0.332	0.498	14.99	0.403	0.446
+ DIFIX3D [35]	14.90	0.460	0.338	16.90	0.527	0.266	17.86	0.560	0.235
+ 3DGS-Enhancer [20]	14.33	0.424	0.464	16.94	0.565	0.356	18.50	**0.630**	0.305
+ DIFIX3D++ [28]	15.93	0.528	0.361	17.95	**0.589**	0.291	18.93	0.619	0.262
+ SetDiff	**16.87**	**0.533**	**0.335**	**18.74**	0.587	**0.264**	**19.60**	0.614	**0.232**

Table 4: **SOTA comparisons on DL3DV** on few-shot 3D reconstruction following [20]. Comparisons of neural reconstruction methods (top) and diffusion-based enhancers (bottom) are reported separately.



Figure 8: **Novel-view images on DL3DV** rendered from sparse-view 3DGS.

Table 5: Ablating the impact of geometric grounding on enhancement quality.

Geometric Grounding	PSNR	SSIM	LPIPS
None	17.59	0.534	0.278
Coordinate maps	18.39	0.557	0.282
Camera poses	18.43	0.560	0.288
Both	**18.74**	**0.587**	**0.265**

Table 6: Ablating the impact of set diffusion on enhancement quality.

$\mathcal{I}^r + \mathcal{I}^t$	1 + 1	2 + 1	3 + 1	3 + 2	3 + 4	3 + 6
PSNR	17.80	18.57	18.67	18.69	18.72	**18.79**
SSIM	0.552	0.578	0.581	0.582	0.583	**0.590**
LPIPS	0.280	0.258	**0.255**	0.255	0.255	0.256

Set diffusion We evaluate the impact of using set-based reference and target inputs, denoted by $\mathcal{I}^r$ and $\mathcal{I}^t$, on enhancement quality. Starting from single-frame enhancement ($|\mathcal{I}^r| = |\mathcal{I}^t| = 1$), we progressively increase the sizes of the reference and target sets and analyze their effect on diffusion-based enhancement performance, as reported in Table 6. The results show that enlarging the reference set leads to a substantial performance gain, as it enables the model to aggregate missing information across multiple views—an essential factor for accurate image restoration. Performance gains begin to saturate beyond three reference views, which appears sufficient to cover the DL3DV dataset. Increasing the size of the target set also improves enhancement quality, albeit to a lesser extent, by exploiting cross-frame correlations.

5 Conclusion

We presented SetDiff, a geometry-grounded, set-based diffusion framework for enhancing novel-view renderings in 3D Gaussian Splatting. By conditioning the denoiser on multiple reference views, pixel-aligned geometric maps, and camera-pose embeddings, SetDiff improves photometric fidelity and suppresses artifacts under sparse observations and wide-baseline extrapolation. Its permutation-equivariant set formulation jointly processes multiple reference and target views with minimal computational overhead. Across four challenging datasets, SetDiff outperforms diffusion-based enhancers and geometry-regularized baselines, making it a robust and scalable component for multi-camera sensor simulation in closed-loop autonomous driving systems.

References

[1] Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)

[2] Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5835–5844. IEEE (2021)

[3] Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19640–19648. IEEE (2023)

[4] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 11618–11628. IEEE (2020)

[5] Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., Zhang, G.: Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. IEEE Transactions on Visualization and Computer Graphics **31**(9), 6100–6111 (2024)

[6] Chen, Z., Yang, J., Huang, J., De Lutio, R., Martinez Esturo, J., Ivanovic, B., Litany, O., Gojcic, Z., Fidler, S., Pavone, M., et al.: Omnire: Omni urban scene reconstruction. In: International Conference on Learning Representations. vol. 2025, pp. 85508–85527 (2025)

[7] Cheng, K., Long, X., Yang, K., Yao, Y., Yin, W., Ma, Y., Wang, W., Chen, X.: Gaussianpro: 3d gaussian splatting with progressive propagation. In: Forty-first international conference on machine learning (2024)

[8] Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28135–28144. IEEE (2025)

[9] Han, X., Jia, Z., Li, B., Wang, Y., Ivanovic, B., You, Y., Liu, L., Wang, Y., Pavone, M., Feng, C., et al.: Extrapolated urban view synthesis benchmark. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28718–28728. IEEE (2025)

[10] Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)

[11] Hwang, S., Kim, M.J., Kang, T., Kang, J., Choo, J.: Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)

[12] Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5885–5894 (2021)

[13] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) (2023)

[14] Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)

[15] Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 170–185 (2018)

[16] Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20775–20785. IEEE (2024)

[17] Li, Y., Wang, Z., Wang, Y., Yu, Z., Gojcic, Z., Pavone, M., Feng, C., Alvarez, J.M.: Memorize what matters: Emergent scene decomposition from multitraverse. Advances in Neural Information Processing Systems **37**, 108389–108438 (2024)

[18] Li, Y., Xiong, K., Guo, X., Li, F., Yan, S., Xu, G., Zhou, L., Chen, L., Sun, H., Wang, B., et al.: Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. arXiv preprint arXiv:2506.08052 (2025)

[19] Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)

[20] Liu, X., Zhou, C., Huang, S.: 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. Advances in Neural Information Processing Systems **37**, 133305–133327 (2024)

[21] Ljungbergh, W., Tonderski, A., Johnander, J., Caesar, H., Åström, K., Felsberg, M., Petersson, C.: Neuroncap: Photorealistic closed-loop safety testing for autonomous driving. In: ECCV (2024)

[22] Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20654–20664 (2024)

[23] Mildenhall, B., Tancik, M., Srinivasan, P.P., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)

[24] Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. In: SIGGRAPH (2022)

[25] Ni, Z., Du, S., Hou, Z., Wu, C., Yang, S.: Para-lane: Multi-lane dataset registering parallel scans for benchmarking novel view synthesis. arXiv preprint arXiv:2502.15635 (2025)

[26] Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5470–5480. IEEE (2022)

[27] NVIDIA, Cao, Y., de Lutio, R., Fidler, S., Cobo, G.G., Gojcic, Z., Igl, M., Ivanovic, B., Karkus, P., Esturo, J.M., Pavone, M., Smith, A., Tanimura, E., Tyszkiewicz, M., Watson, M., Wu, Q., Zhang, L.: Alpasim: A modular, lightweight, and data-driven research simulator for autonomous driving (October 2025), `https://github.com/NVlabs/alpasim`

[28] Omran, M., Zanjani, F., Abati, D., Petersen, J., Habibian, A.: Hybrid gaussian splatting for novel urban view synthesis. arXiv preprint arXiv:2510.12308 (2025)

[29] Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024)

[30] Renz, K., Chen, L., Arani, E., Sinavski, O.: Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In: CVPR (2025)

[31] Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)

[32] Sadat, S., Hilliges, O., Weber, R.M.: Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In: The Thirteenth International Conference on Learning Representations (2024)

[33] Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022), `https://arxiv.org/abs/2202.00512`

[34] Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)

[35] Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models. CVPR (2025)

[36] Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21551–21561 (2024)

[37] Wu, Z., Liu, T., Luo, L., Zhong, Z., Chen, J., Xiao, H., Hou, C., Lou, H., Chen, Y., Yang, R., et al.: Mars: An instance-aware, modular and realistic simulator for autonomous driving. In: CAAI International Conference on Artificial Intelligence. pp. 3–15. Springer (2023)

[38] Xu, Y., Tan, H., Luan, F., Bi, S., Wang, P., Li, J., Shi, Z., Sunkavalli, K., Wetzstein, G., Xu, Z., et al.: Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. arXiv preprint arXiv:2311.09217 (2023)

[39] Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., Peng, S.: Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In: European Conference on Computer Vision. pp. 156–173. Springer (2024)

[40] Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8254–8263. IEEE (2023)

[41] Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.C., Yang, A.J., Urtasun, R.: Unisim: A neural closed-loop sensor simulator. In: CVPR (2023)

[42] Zanjani, F.G., Abati, D., Wiggers, A., Kalatzis, D., Petersen, J., Cai, H., Habibian, A.: Gaussian splatting is an effective data generator for 3d object detection. arXiv preprint arXiv:2504.16740 (2025)

[43] Zhou, H., Lin, L., Wang, J., Lu, Y., Bai, D., Liu, B., Wang, Y., Geiger, A., Liao, Y.: Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. arXiv preprint arXiv:2412.01718 (2024)

[44] Zhou, H., Lin, L., Wang, J., Lu, Y., Bai, D., Liu, B., Wang, Y., Geiger, A., Liao, Y.: Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. IEEE TPAMI (2025)

[45] Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21676–21685 (2024)

[46] Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.H.: Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21634–21643 (2024)

Enhancing Novel View Synthesis via Geometry Grounded Set Diffusion

— Supplementary Material —

A.1 Datasets and Experimental Setup

To comprehensively assess the extrapolation capability of novel view synthesis (NVS) methods under challenging conditions—including lateral camera translation, camera rotation, combined translation and rotation, temporal extrapolation, camera-view sparsity, dynamic scene variation, and both daytime and nighttime environments—we conduct experiments across four diverse datasets. Below, we detail the experimental setup for each dataset.

A.1.1 EUVS Dataset

The Extrapolated Urban View Synthesis (EUVS) dataset [9] is a large-scale autonomous driving benchmark designed to evaluate NVS performance under significant viewpoint shifts, representative of vision-centric autonomous driving systems. It contains 104 real-world urban scenes captured across multiple vehicles, routes, and camera configurations, along with official train/test splits. EUVS specifies three evaluation protocols: (1) Translation-only: cross-lane lateral displacement; (2) Rotation-only: holding out side-view cameras for testing; (3) Translation + rotation e.g., driving along different routes through intersections (considered the most challenging setting).

Training data generation for diffusion enhancers. Each EUVS scene contains multiple recorded traversals, enabling paired supervision for diffusion enhancement. We fit a 3DGS model on one traversal and render it at the camera poses of another recorded traversal. The rendered image serves as the degraded input, and the real image captured at the same pose serves as the clean target. Thus, the reconstruction loss is computed only at observed camera poses with available ground-truth images, not at unobserved synthetic lateral shifts. We repeat this procedure over trajectory combinations within each scene. Para-Lane follows the same protocol using its parallel recorded trajectories.

Test data. To ensure fair evaluation, we adopt a strict scene-level 60%/40% train/test split, preventing overlap or near-duplicate content across sets. For the held-out scenes, we follow the official EUVS protocols: 3DGS is fitted on the designated training trajectory/views, and SetDiff and all baselines are evaluated on the prescribed held-out target views. Accordingly, all reported losses and metrics are computed only against captured images at the corresponding target poses.

Dynamic region handling. Scenes contain multiple asynchronous traversals, causing dynamic objects to differ between training and inference views. Consequently, consistent reconstruction of dynamic regions is not feasible for 3DGS. Following [9], dynamic regions are excluded during both training and evaluation using the dynamic-region masks provided in the dataset.

A.1.2 Para-Lane Dataset

The Para-Lane dataset [25] provides multi-lane urban driving sequences captured from parallel trajectories, enabling evaluation of cross-lane NVS under realistic lane shifts. It includes three cameras and over 16K front-view and 64K surround-view images recorded from synchronized multi-pass runs. As with EUVS, we adopt a strict scene-level 60%/40% split for training diffusion-enhancement models.

For 3DGS reconstruction, we fit one trajectory at a time and treat the remaining parallel trajectories as test sequences—each requiring view-specific enhancement at its corresponding extrapolated viewpoints.



Figure 9: **Multi-camera enhancement** of extrapolated views from nuScenes.

A.1.3 nuScenes Dataset

We use nuScenes [4] to evaluate performance in the presence of dynamic objects, diverse weather, and day/night conditions. The dataset contains 750 urban driving sequences captured by six synchronized cameras.

Since each scene has only a single traversal, we construct a temporal-extrapolation benchmark. Each sequence is split into two disjoint segments: the first 10 seconds are used to fit a dynamic 3DGS model with OmniRe, and the following 5 seconds are used as novel future views. We group these future frames into five timestamp buckets to evaluate increasing extrapolation distances. The final bucket, corresponding to 14–15 seconds, is the most challenging due to accumulated 3DGS drift, dynamic object motion, and scene incompleteness. Dynamic objects are retained during both training and evaluation.

For diffusion-enhancer training, we use a scene-level split with 600 training scenes and 150 test scenes; after sorting scene names, every fifth scene is assigned to the test set. For each nuScenes packet, the six synchronized camera views from one timestamp are used as references, and the six corresponding views from a future timestamp bucket are used as targets. This enables SetDiff to enhance the full surround-view rig jointly, rather than processing each camera independently.

Joint multi-camera enhancement is important because reference-target overlap varies across front, side, and rear cameras under future ego-motion, and dynamic objects create camera-dependent occlusions. The set mixer allows information sharing across overlapping or adjacent camera views, while geometry grounding anchors each pixel to the rendered 3D structure. This reduces independent per-camera hallucinations and helps preserve the global scene layout. Figure 9 shows qualitative multi-camera examples, and additional videos are provided on the project page.

A.1.4 DL3DV Dataset

To assess performance beyond autonomous driving, we additionally evaluate on the DL3DV-10K dataset [19], which spans a broad spectrum of bounded and unbounded indoor and outdoor scenes.

This enables direct comparison with 3DGS-Enhancer [20] under varying input sparsity levels (3, 6, and 9 training views). We follow the original splitting strategy from [20], comprising 130 training scenes and 20 test scenes.

For each sparsity level, we adopt the exact train/test camera selections defined in [20]. A 3DGS model is fitted using only the sparse training views, and rendered images at the held-out camera poses are used to evaluate enhancement quality.

A.2 Implementation Details

Training Configuration. The proposed SetDiff framework is trained on packets consisting of $N + M$ frames, where N denotes the number of reference views and M denotes the number of target views. For the EUVS, Para-Lane, and DL3DV datasets, the packet length is fixed to 8. For nuScenes, which contains 6 cameras, we use a packet length of 12 (i.e., 6 reference and 6 target views). During inference, the packet length is increased to 16 for EUVS, Para-Lane, and DL3DV, and to 24 for nuScenes.

Training is conducted on four NVIDIA A100-80G GPUs with a total batch size of 4, using approximately 40K iterations. For classifier-free guidance (CFG), we apply a dropout rate of 15% to both the C-map and camera pose inputs. At inference time, the CFG scale is set to 2.0. All experiments are performed at a resolution of 576×1024 pixels. Following DiFix3D [35], we fine-tune SD-Turbo using a reduced noise level ($\tau = 200$ instead of $\tau = 1000$), rather than introducing random Gaussian noise.

3D Priors. For each packet containing reference and target frames, camera poses and their corresponding C-maps are transformed from world coordinates into the coordinate frame of the first target camera. Prior to computing Plücker ray embeddings, we normalize the camera poses. This is essential because the cross-product term $\mathbf{o} \times \mathbf{d}$ is not scale-invariant, as it directly depends on the magnitude of $\mathbf{o}$. Moreover, camera poses estimated by Structure-from-Motion (SfM) often exhibit scale inconsistencies across scenes and datasets.

To address these issues, we normalize all camera positions during both training and inference. Specifically, the positions are scaled such that the maximum pairwise distance among cameras within a packet is equal to unity. This normalization alleviates scale ambiguity, stabilizes training, and ensures consistent behavior across diverse scene configurations.

A.2.1 Additional Ablations

Architecture	PSNR	SSIM	LPIPS
w/o LoRA	18.604	0.584	0.287
cross-attn	18.673	0.587	0.268
Ours	**18.74**	**0.587**	**0.265**

Table 7: Effect of architectural choices on enhancement quality.

We evaluate two architectural variants: (i) removing the LoRA-based fine-tuning of the VAE decoder (*"w/o LoRA"*), and (ii) replacing additive conditioning with cross-attention in the 2D UNet (*"cross-attn"*). As shown in Table 7, LoRA fine-tuning yields a modest improvement in LPIPS, suggesting enhanced perceptual fidelity and realism. The cross-attention variant performs comparably to the baseline, indicating that it provides a viable alternative mechanism for injecting latent priors into the UNet.

A.2.2 Consistency Evaluation

We evaluate view consistency on nuScenes using the Warp error metric [15]. As shown in Table 8, SetDiff reduces warp error compared to both OmniRe and DiFix++, while substantially improving PSNR, SSIM, and LPIPS. This suggests that the proposed enhancement does not improve reconstruction metrics merely by smoothing the outputs, but also preserves, and slightly improves, view-to-view alignment.

	Consistency↓	PSNR	SSIM	LPIPS
OmniRe	111.0	17.15	0.645	0.448
+ DiFix++	110.2	18.74	0.696	0.375
+ SetDiff	**108.9**	**21.58**	**0.716**	**0.303**

Table 8: Evaluating the enhancement consistency.

SetDiff is designed as a view-set enhancer rather than a replacement for geometry-based rendering. The underlying 3DGS or OmniRe renderer remains responsible for the global scene geometry, while SetDiff corrects local rendering artifacts caused by under-reconstruction, sparse observations, or extrapolated viewpoints. Since the correction is performed in image space, it does not provide the hard multi-view consistency guarantees of explicit 3D rendering. To mitigate this limitation, each target view is conditioned on geometry rendered from the same 3D representation, camera-pose embeddings, and multiple nearby clean reference views. In addition, the set mixer enables target views within the same packet to exchange information during denoising, reducing independent frame-wise artifacts compared to single-view enhancement.

The marginal improvement in warp error is expected: geometry-based renderers already impose strong cross-view constraints, leaving less room for improvement in consistency than in image quality. Nevertheless, the results indicate that SetDiff improves perceptual and reconstruction quality without introducing additional temporal or cross-view drift. Residual local inconsistencies may still occur under extreme occlusion or very small reference-target overlap, as discussed in Sec. A.4.

A.3 Runtime Analysis and Memory Consumption

To quantify the scalability of SetDiff, we measure inference time and peak GPU memory usage while varying the number of conditioning (reference) views. Here, we set the number of target views to one. In practice, we usually enhance more than one target view at a time to perform joint enhancement, for example, in a multi-camera setup. Nevertheless, the reported measurements remain valid for the evaluated configuration.

# Reference Views	Comp. Time (ms)	GPU Memory (GB)
1	310	4.82
5	832	10.44
10	1,508	15.01
15	2,276	21.28
20	3,101	27.57
23	3,570	31.34

Table 9: Runtime and peak GPU memory of SetDiff for different numbers of reference views.

Both inference time and memory usage scale approximately linearly with the packet size over this range, increasing from approximately 0.3 s / 4.8 GB for 1 reference view to 3.6 s / 31.3 GB for 23 reference views. Although SetDiff jointly processes all views through the Set Mixer, the measured runtime and memory growth are close to linear with respect to the number of input views in the examined range. This suggests that the framework can effectively leverage additional references while maintaining predictable computational requirements. In practice, we usually enhance more than one target frame at a time, so the decoder processing time slightly increases, depending on the number of target views.

The reported numbers correspond to enhancement throughput, not raw 3DGS rendering speed. In our pipeline, the 3D renderer first produces the initial novel views, after which SetDiff applies post-rendering correction. Given its inference latency, SetDiff targets offline or near-online applications, including simulation, dataset augmentation, closed-loop evaluation, and reinforcement-learning-based VLA training. In these settings, the visual quality gains from jointly processing multiple views can justify the additional computation and latency.

A.4 Failure cases

SetDiff leverages visual information from clean reference views to guide enhancement when sufficient geometric overlap exists between reference and target views. When this overlap becomes very

limited—due to extreme viewpoint shifts or heavy occlusions—the geometric conditioning becomes less reliable. In such cases, the model relies more heavily on its diffusion prior, yielding visually plausible results but potentially introducing misalignment or localized hallucinations. These situations are uncommon in autonomous driving scenarios, where novel views typically share substantial visible scene content with the recorded frames. Nonetheless, improving robustness under minimal geometric overlap remains an important direction for future work, for example through uncertainty-aware conditioning or adaptive weighting of geometric cues.

Moreover, we observed occasional photometric saturation artifacts, which are primarily caused by the classifier-free guidance (CFG) scale used during inference. In the nuScenes experiments, the input renderings are often severely degraded due to extreme viewpoint extrapolation. To strengthen the conditioning signal and improve restoration quality, we employed a relatively large CFG scale. Although this generally enhances fidelity and scene consistency, it can also amplify undesirable color saturation artifacts. The relationship between high CFG scales and oversaturated colors, overexposed outputs, and other visual artifacts has been widely discussed in the diffusion-model literature [32]. Excessive guidance can push generated samples away from the training distribution, leading to degraded visual quality and saturation effects. Our empirical observations are consistent with these findings: saturation artifacts occur primarily in the most challenging extrapolation scenarios, where stronger guidance is required to recover substantial missing content. We therefore attribute this behavior to the current inference configuration, rather than to a limitation of the SetDiff architecture itself.

A.5 Qualitative Video Results

For qualitative visualizations of enhanced test videos, please refer to the project webpage provided in HTML format.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are supported by the proposed SetDiff architecture and the empirical evaluation. The paper demonstrates geometry-grounded, set-based diffusion enhancement for variable-size multi-view inputs, and validates the claimed improvements in photometric fidelity, perceptual quality, and robustness on four datasets: EUVS, Para-Lane, nuScenes, and DL3DV.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses limitations and failure cases in the supplementary material, including reduced reliability under extreme viewpoint shifts, heavy occlusion, or very limited reference-target overlap, where the model may rely more strongly on the diffusion prior and produce localized hallucinations or misalignment. Runtime and deployment constraints are also discussed in the runtime analysis.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A].

   Justification: The paper does not present theoretical results, theorems, or formal proofs. Its contributions are methodological and empirical, focusing on a diffusion-based multi-view enhancement framework and experimental validation.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes].

   Justification: The paper provides the key information needed to reproduce the main experimental results, including dataset construction and train/test splits, model architecture, geometric conditioning, packet configuration, training resolution, optimizer-related training setup, number of iterations, CFG settings, and compute resources. Additional implementation and dataset details are provided in the supplementary material.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No].

   Justification: The experiments use publicly available datasets and publicly available backbone components, including the initialized SD-Turbo checkpoint. The paper describes the experimental setup. The code will be released publicly upon acceptance, facilitates reproduction of the main results.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: The paper specifies the main experimental settings, including dataset splits, packet lengths for different datasets, training and inference configurations, image resolution, number of iterations, GPU setup, classifier-free guidance dropout and inference scale, and the use of SD-Turbo fine-tuning. These details are given in the main paper and supplementary implementation section.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No].

   Justification: Training the diffusion models, along with competitive methods on several experimental setups and multiple large-scale datasets is impractical and not common on such a target task.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes].

   Justification: In section A.3 of the supplementary materials we discussed and reported the compute setup used for training, including four NVIDIA A100-80GB GPUs, batch size, number of iterations, and training resolution. The supplementary runtime analysis also reports per-frame enhancement time and FPS on an A100-80GB GPU, and discusses the intended offline or near-online simulation use case.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

Justification: The authors reviewed the NeurIPS Code of Ethics and the work is conducted as an empirical study on novel-view enhancement using existing datasets and models. The paper does not involve human-subject studies, crowdsourcing, or the release of high-risk personally sensitive data.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: Although the paper does not directly propose a full autonomous-driving simulator, it discusses use cases in visual-data augmentation, simulation, and closed-loop evaluation. Potential positive impacts include improved simulation fidelity and safer evaluation in controlled environments. Potential negative impacts as stated in the failure cases section include over-reliance on enhanced visual outputs and possible hallucinations or misalignments under extreme viewpoint shifts or occlusions, especially if used without validation in safety-critical settings.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: Although the method fine-tunes a diffusion-based image generator, it is not released or used as an open-ended image-generation system. The model is conditioned on rendered novel views, geometry maps, camera poses, and reference images for the specific task of post-rendering enhancement. Therefore, the work does not introduce the same misuse risks as general-purpose text-to-image or unrestricted image-generation models.

Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes].

    Justification: The paper cites the existing datasets, models, and prior methods used in the experiments, including public datasets and the SD-Turbo backbone.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, benchmark, or standalone asset. It evaluates the proposed method using existing datasets and existing rendering/diffusion components.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

Answer: [N/A]

Justification: The work does not involve crowdsourcing, human-subject experiments, user studies, or participant-provided annotations. Therefore, participant instructions, screenshots, and compensation details are not applicable.

Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

Answer: [N/A]

Justification: The paper does not involve research with human participants, crowdsourced workers, or user studies. Consequently, IRB approval or equivalent human-subjects review is not applicable.

Guidelines:

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

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The work does not include use of LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.