SetDiff: Enhancing Novel View Synthesis via Geometry Grounded Set Diffusion

Farhad G. Zanjani Hong Cai Amirhossein Habibian

Qualcomm AI Research

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

The problem: 3DGS breaks off the training trajectory



3DGS artifacts grow as the camera moves away from the training trajectory (less than one lane, one lane, two lanes); SetDiff enhances them.

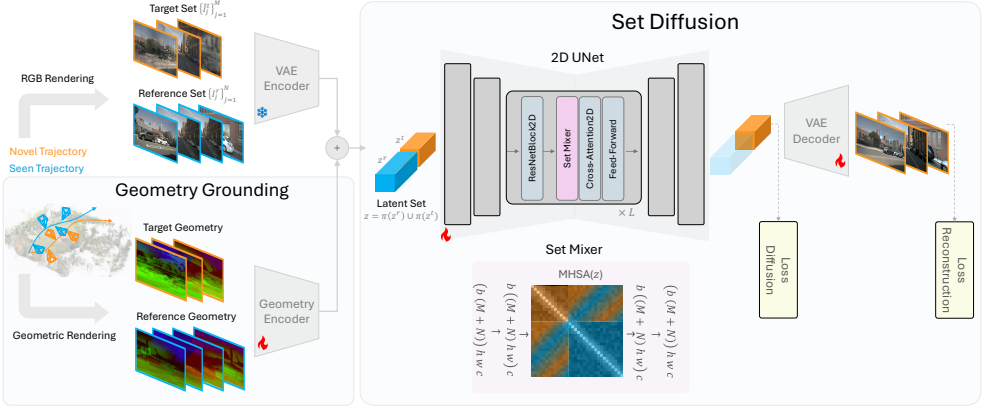
Why it matters

- 3DGS is a strong backbone for sensor simulation in autonomous driving
- Closed-loop simulation constantly leaves the training trajectory

Existing diffusion enhancers

- Condition only on RGB: no explicit 3D guidance
- Enhance novel-view frames by conditioning mostly on a single reference
- Hallucinate content that drifts from the true scene, especially at low SNR
- Do not fully leverage the visual content of the observed images

SetDiff overview framework



Geometry-grounded: conditions on rendered 3D dense correspondences (C-maps) and camera poses

Set diffusion: jointly enhances a variable-size set of M target views from N references

Efficient: single-step distilled latent diffusion (SD-Turbo)

Geometric grounding: C-maps and camera poses

3D coordinate maps (C-maps)

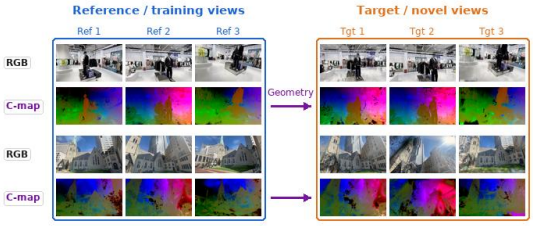
Opacity-weighted accumulation of Gaussian centres x_i along each ray gives a pixel-aligned cross-view 3D correspondences:

$$c(u, v) = \sum_i \alpha_i(u, v) \prod_{j<i} (1 - \alpha_j(u, v))\, x_i$$

- α_i: Gaussian opacity
- Like positional encodings, C-maps guide attention across unordered views, even when noisy

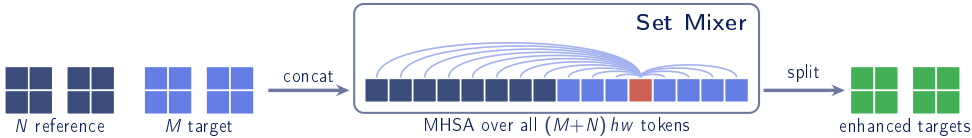
Camera poses

- Encoded as Plücker rays with normalized extrinsics
- A light-weight encoder Ψ embeds C-maps and rays; the result is added to the VAE latents



Rendered occlusion-aware C-maps for two DL3DV scenes: colors encode 3D correspondences between reference and target views, a useful geometric guide even when noisy.

Set diffusion: Set Mixer and training



Set Mixer

- U-Net self-attention layers are replaced by MHSA over the union of all tokens of the $M+N$ images ($b \times (M+N)hw \times c$)
- No temporal layers or position encoding $\Rightarrow$ any number and order of views

Training

- Fine-tune SD-Turbo U-Net with single-step v-prediction; VAE decoder (LoRA) and geometry encoder are tuned jointly
- Selective decoding: only a random subset of target latents is decoded per step, keeping the pixel-space loss affordable

$$\mathcal{L} = \mathbb{E}\left[\|\mathrm{y}^t - f_\theta([\mathrm{z}^r; \mathrm{z}^t], \tau)\|_2^2\right] + \mathcal{L}_{\mathrm{Rec}}\,(\ell_2 + \mathrm{SSIM} + \mathrm{LPIPS})$$

(y^t: clean target, z^t: noisy target, z^r: clean references, τ: noise level)

Quantitative results

Para-Lane

	Single lane			Double lane		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
3DGS	17.05	0.524	0.446	16.26	0.505	0.472
+ DiFix3D	17.62	0.470	0.234	16.87	0.447	0.271
+ DiFix3D++	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**

nuScenes (extrapolated)

	PSNR	SSIM	LPIPS
OmniRe	17.15	0.645	0.448
+ DiFix3D	17.12	0.612	0.371
+ DiFix3D++	18.74	0.696	0.375
+ **SetDiff**	**21.58**	**0.716**	**0.303**

DL3DV (PSNR / LPIPS)

	Sparse Recon.		
	3 views	6 views	9 views
3DGS	10.97/.567	13.34/.498	14.99/.446
+ DiFix3D	14.90/.338	16.90/.266	17.86/.235
+ DiFix3D++	15.93/.361	17.95/.291	18.93/.262
+ **SetDiff**	**16.87/.335**	**18.74/.264**	**19.60/.232**

- **SetDiff outperforms competitive diffusion enhancers** across extrapolated NVS setups and datasets
- Inference latency grows **approximately linearly** with the number of reference views (0.3 s for 1 view to 3.6 s for 23)

Qualitative results and takeaways



Novel views on nuScenes. Odd rows: reference and extrapolated OmniRe renderings; even rows: SetDiff enhancements.

- Geometry-grounded conditioning (C-maps + poses) reduces hallucination and improves fidelity
- Set diffusion jointly enhances variable-size sets of views with a single-step model
- State of the art on EUVS, Para-Lane, nuScenes and DL3DV
- Enables closed-loop simulation for autonomous driving and robotics

Thank you!

Project page

qualcomm-ai-research.github.io/setdiff

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