

Scene Splatter: Momentum 3D Scene Generation from Single Image with Video Diffusion Model

Supplementary Material

1. Additional Implementation Details

Flash3D [2] predicts 2 Gaussians for each pixel. The comprehensive configuration for Gaussian optimization is shown in Table 1.

Table 1. Implementation details of Gaussian Optimization.

| Config | Parameter |
|-----------------------------------|-----------|
| sh degree | 3 |
| initialize position learning rate | 0.00003 |
| feature learning rate | 0.001 |
| opacity learning rate | 0.01 |
| scaling learning rate | 0.0002 |
| rotation learning rate | 0.0002 |
| densification interval | 100 |
| densify gradient threshold | 0.0002 |

2. Additional Ablation Studies

We conduct ablation studies to investigate our momentum coefficients. We replace our designed latent-level momentum coefficient in Eq. 8 by fixed values, where $\lambda = 0.3, 0.5, 0.7, 0.9$. We report the quantitative results in Table 2 and visualize the rendering results in Figure 1. Without our adaptive λ_i^z defined in Eq. 10, the video diffusion model can not recover the distortions.

Table 2. Ablation study of our latent momentum coefficient. We report the average PSNR, SSIM and LPIPS of rendering results.

| Method | PSNR \uparrow | SSIM \uparrow | LPIPS \downarrow |
|-----------------|-----------------|-----------------|--------------------|
| Flash3D [2] | 15.87 | 0.640 | 0.349 |
| Ours | 17.58 | 0.703 | 0.268 |
| $\lambda = 0.3$ | 17.51 | 0.694 | 0.269 |
| $\lambda = 0.5$ | 17.58 | 0.692 | 0.269 |
| $\lambda = 0.7$ | 17.57 | 0.693 | 0.270 |
| $\lambda = 0.9$ | 17.55 | 0.689 | 0.268 |

Besides, we also replace our designed pixel-level momentum coefficient in Eq. 13 by fixed values, where $\mu = 0.3, 0.5, 0.7, 0.9$. We report the quantitative results in Table 3 and visualize the rendering results in Figure 2. As shown in Figure 2, other coefficients suffer from artifacts to merge $\Phi_\lambda(\mathcal{I})$ and $\Phi_0(\mathcal{I})$, where our method can achieve reasonable results by balancing the consistency of scenes

Table 3. Ablation study of our pixel momentum coefficient. We report the average PSNR, SSIM and LPIPS of rendering results.

| Method | PSNR \uparrow | SSIM \uparrow | LPIPS \downarrow |
|-------------|-----------------|-----------------|--------------------|
| Flash3D [2] | 15.87 | 0.640 | 0.349 |
| Ours | 17.58 | 0.703 | 0.268 |
| $\mu = 0.3$ | 17.18 | 0.695 | 0.292 |
| $\mu = 0.5$ | 17.29 | 0.697 | 0.283 |
| $\mu = 0.7$ | 13.74 | 0.660 | 0.306 |
| $\mu = 0.9$ | 15.42 | 0.668 | 0.317 |

and the generative power of video diffusion models. Results in Table 3 also demonstrate that our method outperforms the fixed pixel-level momentum coefficient in all matrices.

3. Analysis on Camera Trajectories

To generate a video with M frames, the total iteration is $h = \lceil \frac{M-N}{N-n} \rceil + 1$. A larger n indicates more time and error accumulation due to more iterations, while a smaller n leads to more inconsistency with less reference information. We provide more results with different camera trajectories in Table 4 and Figure 3, which demonstrate the generalization of our method. We select n as a trade-off between efficiency and performance, and it can be set to different values for requirements.

Table 4. Analysis of n with different camera trajectories. We report iterations for 100 frames.

| n | Iter | In | Out | Rotate | Up | Down |
|-----|------|-------|-------|--------|-------|-------|
| 0 | 4 | 22.72 | 19.11 | 20.06 | 15.33 | 21.33 |
| 5 | 5 | 22.86 | 19.20 | 20.10 | 15.35 | 21.59 |
| 10 | 6 | 22.73 | 19.12 | 20.20 | 15.36 | 21.74 |
| 15 | 9 | 22.12 | 18.85 | 20.18 | 15.29 | 21.68 |

4. More comparison

We follow the settings in ReconFusion and provide more comparison with sparse view methods in Table 5. Our method consumes less time and achieves better SSIM and LPIPS results than ZeroNVS even with 3 input views on Re10K. Each iteration (Diff. ~ 2 min, Recon. ~ 1.5 min) requires similar inference time to MVSp360. For further comparison of our method and baselines [2-4], we provide videos of the rendering results as attachments.

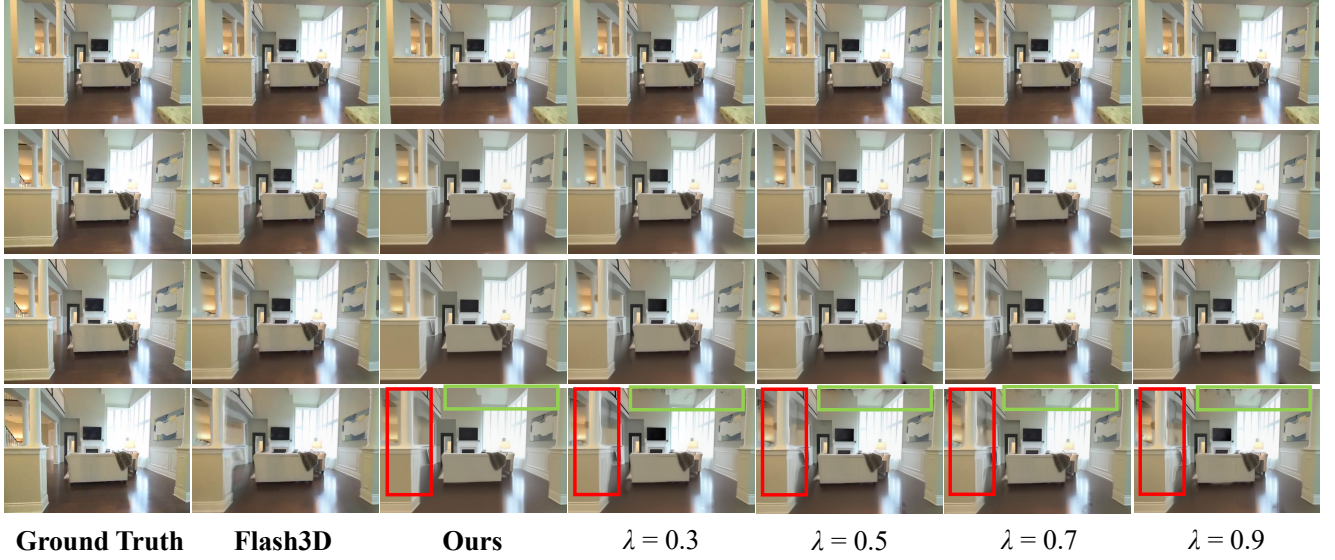


Figure 1. Visualization of additional ablation study on latent-level momentum coefficient.

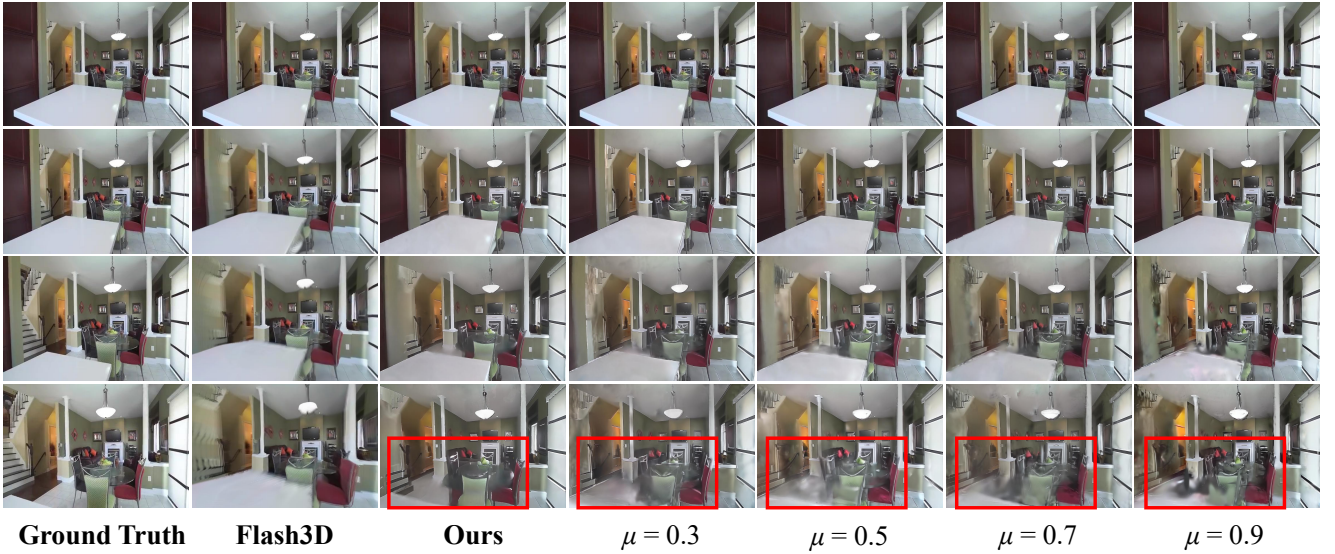


Figure 2. Visualization of additional ablation study on pixel-level momentum coefficient.

References

- [1] Yuedong Chen, Haoifei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, pages 370–386. Springer, 2024. [3](#)
- [2] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *NeurIPS*, 2024. [1](#)
- [3] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- [4] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. [1](#), [3](#)

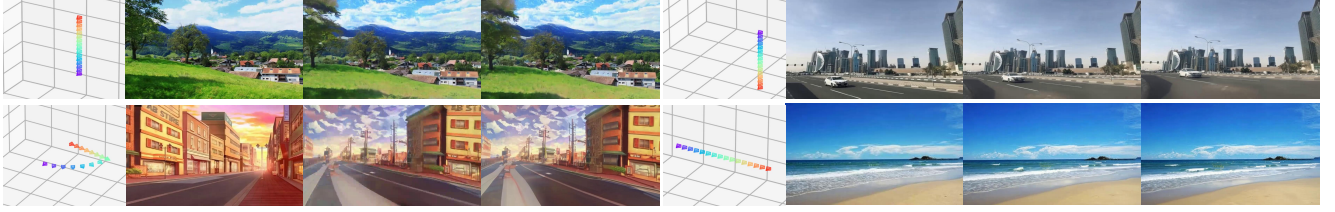


Figure 3. Results of our methods on more camera trajectories.

Table 5. More quantitative comparison with other methods. † Our method contains only one iteration with interval sampling.

| Method | View | RealEstate10K | | | MipNeRF360 | | | Time |
|-------------------|------|---------------|-------|--------|------------|-------|--------|---------|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | |
| ViewCrafter [4] | 1 | 13.72 | 0.450 | 0.547 | 11.77 | 0.297 | 0.754 | 13 min |
| Ours | 1 | 17.04 | 0.680 | 0.287 | 12.00 | 0.315 | 0.750 | 17 min |
| Ours [†] | 1 | 16.77 | 0.680 | 0.287 | 11.95 | 0.325 | 0.750 | 3.5 min |
| MVSplat [1] | 3 | 23.77 | 0.858 | 0.174 | | | | <1s |
| MVSplat360 | 3 | 20.60 | 0.787 | 0.227 | | | | 3 min |
| ZeroNVS | 3 | 19.11 | 0.675 | 0.422 | 14.44 | 0.316 | 0.680 | 60 min |
| ReconFusion | 3 | 25.84 | 0.910 | 0.144 | 15.50 | 0.358 | 0.585 | |
| CAT3D | 3 | 26.78 | 0.917 | 0.132 | 16.62 | 0.377 | 0.515 | |
| ReconX | 3 | | | | 17.16 | 0.435 | 0.407 | |
| 3DGS-Enhancer | 6 | | | | 13.96 | 0.260 | 0.689 | |