# StreetCrafter: Street View Synthesis with Controllable Video Diffusion Models

Supplementary Material

## 1. More Implementation Details

### 1.1. StreetCrafter Training Details

We construct the training video clips using the front camera and LiDAR sensor of Waymo Open [6] and PandaSet [9] datasets, with the start frame of each video clip selected at the interval of 0.5 second (5 frames for both dataset). We set the radius of each LiDAR point cloud in NDC space to 0.01 and crop the upper part of LiDAR condition maps to match the input resolution of the diffusion model during both training and inference.

For adaptation from the pretrained model of Vista [2], we ignore the action control layer injected via cross-attention and mark the first element of the frame-wise mask to 1 and the rest to 0. We incorporate the LoRA [3] adapters introduced during the learning of action controllability as it contributes to the enhancement of visual quality [2]. More details can be found in the original paper.

During the low-resolution training stage, we sample exclusively from Waymo dataset. During the high-resolution training stage, we sample from a hybrid dataset, combining Waymo Open and PandaSet datasets with sampling probabilities of 0.9 and 0.1, respectively.

### 1.2. StreetCrafter Distillation Details

**Loss function.** We jointly optimize the gaussian parameters of background and foreground moving objects, texel of the high-resolution sky cubemap and noisy object tracklets following Street Gaussians [10]. The extra loss $\mathcal{L}_{\text{g}}$ for input view camera is defined as:

$$\mathcal{L}_{\text{g}} = \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{sky}}\mathcal{L}_{\text{sky}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{1}$$

where $\mathcal{L}_{\text{depth}}$, $\mathcal{L}_{\text{sky}}$ and $\mathcal{L}_{\text{reg}}$ share the same format as Street Gaussians [10]. Please refer to the original paper for more details. The coefficients $\lambda_{\text{depth}}$, $\lambda_{\text{sky}}$ and $\lambda_{\text{reg}}$ in Equation 1 are set to 0.01, 0.05 and 0.1, respectively. For the loss function of novel view camera, we crop the upper part of the rendering image and resize to $576 \times 1024$ to compute the LPIPS [15] loss with novel view image generated by StreetCrafter .

**Point cloud initialization.** We initialize the background gaussian model as the combination of LiDAR and SfM point cloud following. The object gaussian model is initialized with aggregated LiDAR points obtained from object tracklets or random sampling. The colors of LiDAR points are assigned by projecting them onto the nearest image plane.

**Optimization.** We adopt the densification strategy introduced in [14] to prevent suboptimal solutions by accumulating the norms of view-space position gradients. The densification threshold is set to 0.0006. We disable the pruning of big gaussians in world space since this hinders the gaussian model to represent distant regions and the LiDAR points have provided a good initialization to prevent the model from falling into local optima. We finally introduce the 2D Mip filter to enable anti-aliased rendering inspired by [13].

We sample StreetCrafter every 5000 iterations from iteration 7000 to 22000 and linearly reduce the noise scale $s$ from 0.7 to 0.3. We use the annotated object tracklets provided by the datasets, with the learning rates for the translation vector and rotation matrix initialized at $5e^{-4}$ and $1e^{-5}$, respectively, decaying exponentially to $1e^{-5}$ and $5e^{-6}$. For the remaining parameters, we use the default values from the official implementation of Street Gaussians [10].

### 1.3. Evaluation Details

**Interpolation.** For the interpolation setting of Ours-V in the main paper, we incorporate training images along the input trajectory in addition to the reference image and LiDAR conditions. During each denoising step, we replace the prediction of $\mathcal{F}_\theta$ at training frame with the clean latent of training images. This could lead to improvement in the interpolation quality, with PSNR increasing from 23.66 to 27.19 and LPIPS decreasing from 0.098 to 0.087 on Waymo Open Dataset [6].

**Baselines.** We use the same object tracklets as our method for all the baselines requiring 3D bounding box as input [7, 10, 11]. We use the same rendering kernel and optimization strategy as our method for all the baselines using 3DGS as the scene representation [5, 10].

**Metric.** For Fréchet inception distance (FID) metric, we mark the input video as real and the rendered sequence as unreal for each scene. For Ours-V, we upper-crop the input video to match the resolution of the generated video.

## 2. Additional Experiments

### 2.1. More Comparisons

**Comparisons with baselines.** We provide more qualitative comparisons on Waymo [6] dataset under the setting of lane change in Figure 6. Figures 7, 8 display the view interpolation results on Waymo Open [6] and PandaSet [9] datasets. Our method achieves comparable rendering

quality to the baselines while achieving significantly better results for view extrapolation.

**Comparisons with concurrent works.** We compare Ours-V with ViewCrafter [12] and Ours-G with DriveDreamer4D [16], both of which are concurrent of our works. For ViewCrafter, we make several modifications to improve its performance. First, we build the global point cloud from the whole input sequence instead of selecting the first frame as in the original setting. Second, we fix the camera parameters during the global alignment process of DUSt3R [8] by using camera calibration results from the dataset. This can also help define camera poses within the dataset's coordinate system when performing view extrapolation. For DriveDreamer4D, we compare our method on segment-103593 of Waymo following their setting with PVG [1] as the base model.

As shown in Table 1 and Figures 1, 2, our method achieves better view synthesis results under both input and novel trajectory compared with ViewCrafter [12]. ViewCrafter builds point cloud without considering that the geometry of dynamic scene changes overtime, thus it fails to accurately model dynamic regions as shown in Figure 2. The generated results also degrade significantly as the camera deviates from the input trajectory since the predicted point cloud is defined in the camera coordinate system. In contrast, the LiDAR point cloud defined in the world coordinate system provides our model with stronger generalization ability on new trajectories even if no ground truth data is available during training.

As shown in Figure 3, our method achieves better results under novel trajectory compared with DriveDreamer4D [16] (FID @ 2m 89.71 vs. 91.23, FID @ 3m 96.11 vs. 123.32). Due to the sparse conditioning of DriveDreamer4D, the generated videos often lack accurate 3D perception. This leads to noticeable artifacts in the reconstructed scene, such as the vehicle in the lower left corner and the building on the right side.

| Methods | Input trajectory | | | Novel trajectory | |
|---|---|---|---|---|---|
| | PSNR↑ | LPIPS↓ | FID ↓ | FID↓ @ 2m | FID↓ @ 3m |
| ViewCrafter [12] | 21.59 | 0.226 | 97.86 | 135.69 | 137.76 |
| Ours-V | **25.90** | **0.143** | **60.49** | **62.43** | **73.49** |

Table 1. Quantitative comparison with ViewCrafter [12]. We use the video clips in ablation to test the results on input trajectory and the scenes in experiment to test the results on novel trajectory. Metrics are averaged over all sampled sequences.

## 2.2. More Editings

We provide more visual results of scene editing in Figure 4 including object translation, replacement and removal.

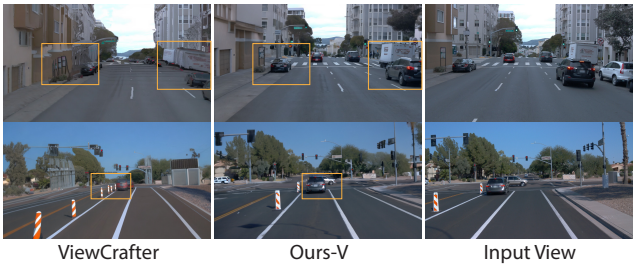


Figure 1. Qualitative comparison with ViewCrafter [12] under novel trajectory. The camera is laterally shifted for 3 meters. Input view denotes the closest input video frames.


Figure 2. Qualitative comparison with ViewCrafter [12] under input trajectory. Our model can handle moving objects.


Figure 3. Qualitative comparison with DriveDreamer4D [16] under novel trajectory. The camera is laterally shifted for 2 meters.

## 2.3. More Ablations

**Analysis of LiDAR conditions** We present more visual comparisons of the design choice of StreetCrafter in Figure 5. The generated frames under the guidance of camera parameter as vector are blurry when the target viewpoint move away from the reference image. Although the 3D bounding box can provide priors regrading object motions, it still fails to align well with the target image as shown in the first row of Figure 5. The results under the condition of projected multi-frame LiDAR can preserve the scene structure but still lack details in regions with rich texture.

**Analysis of the novel view sampling ratio** We conduct experiment on one Waymo sequence to analyze the influence of novel view sampling ratio $p$. The results in Table 2 indicates that $p = 0.4$ yields the overall best result.

| Methods | Interpolation | | Lane Shift | |
|---|---|---|---|---|
| | PSNR↑ | LPIPS↓ | FID↓ @ 2m | FID↓ @ 3m |
| (1) $p = 0.8$ | 28.76 | 0.059 | 72.76 | 84.78 |
| (2) $p = 0.6$ | 29.61 | 0.049 | 68.33 | 81.26 |
| (3) $p = 0.4$ | **30.42** | **0.041** | 67.54 | **79.19** |
| (4) $p = 0.2$ | 30.29 | **0.041** | **67.09** | 81.26 |

Table 2. Ablations on the novel view sampling ratio $p$.


Figure 4. **More editing results on the Waymo [6] dataset.** Images in the right and left columns represent the results before and after editing, respectively.

**Analysis of the noise scale** We conduct experiment on one Waymo sequence to analyze the influence of noise scale $s$.s We have demonstrated in the main paper that adding noise to the render latents leads to better scene consistency than starting from gaussian noise. Since the added noise would have little influence when $s$ is less than 0.3 according to the sampling scheme of Vista [4], we set $s_{\text{min}}$ to 0.3 and ablate on the value of $s_{\text{max}}$. The results in Table 3 indicates that reducing $s$ from 0.7 to 0.3 maintains a balance between sampling steps and rendering quality.

| Methods | Interpolation | | Lane Shift | |
|---|---|---|---|---|
| | PSNR↑ | LPIPS↓ | FID↓ @ 2m | FID↓ @ 3m |
| (1) $s_{\text{max}} = 1.0, s_{\text{min}} = 0.3$ | 30.08 | 0.044 | 69.66 | 79.87 |
| (2) $s_{\text{max}} = 0.7, s_{\text{min}} = 0.3$ | 30.42 | **0.041** | **67.54** | **79.19** |
| (3) $s_{\text{max}} = 0.5, s_{\text{min}} = 0.3$ | **30.46** | 0.042 | 68.68 | 81.23 |

Table 3. Ablations on the noise scale $s$.

## 2.4. Deformable Objects

We show the generated videos of StreetCrafter under scene with multiple deformable objects such as pedestrians in Figure 9. Although multi-frame LiDAR conditions are not ideal for deformable objects, our model can generate plausible results with the generative prior of diffusion model.


Figure 5. **Visual ablation results on the design choice of StreetCrafter.**


Figure 6. **Qualitative comparisons on the Waymo [6] dataset.** The camera is laterally shifted for 2 meters to left or right. Input view refers to the closest training camera.


Figure 7. **Qualitative comparisons of view interpolation on the Waymo [6] dataset**.

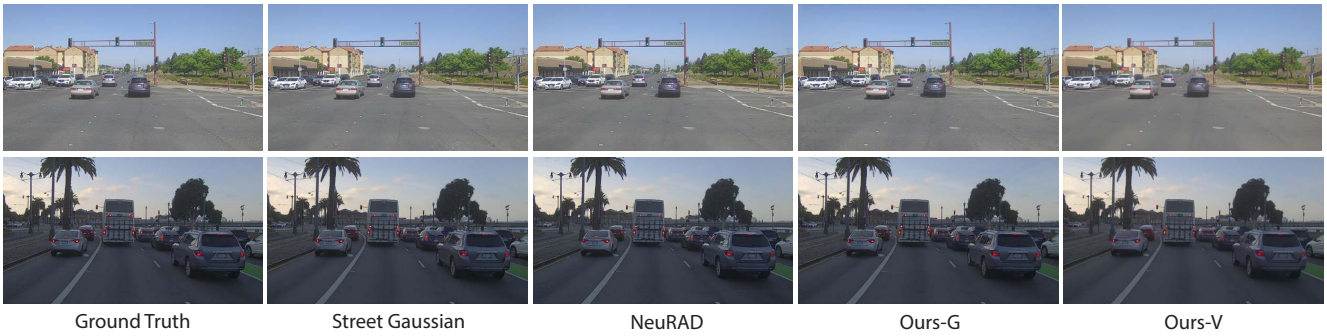


Figure 8. **Qualitative comparisons of view interpolation on the PandaSet [9] dataset**.

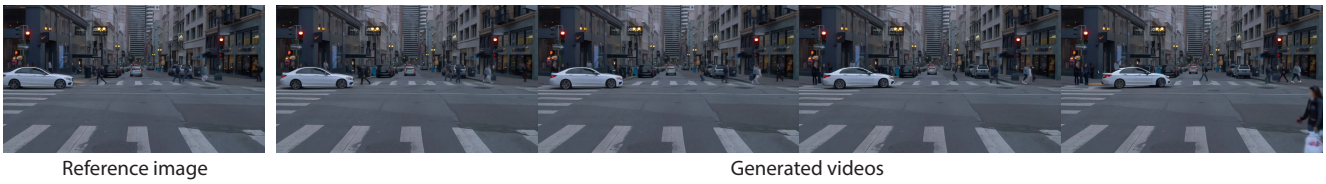


Figure 9. **Visual results of StreetCrafter for scene with deformable objects**.

## References


[1] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv:2311.18561*, 2023. 2

[2] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *NeurIPS*, 2024. 1

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 1

[4] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 3

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

[6] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1, 3, 4, 5

[7] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *CVPR*, 2024. 1

[8] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2

[9] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE, 2021. 1, 5

[10] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *ECCV*, 2024. 1

[11] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *CVPR*, 2023. 1

[12] 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. 2

[13] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, 2024. 1

[14] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics*, 2024. 1

[15] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[16] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. Drivedreamer4d: World models are effective data machines for 4d driving scene representation, 2024. 2