

1 Implementation Details

Implementation details of Geometry Priors. To construct our geometry priors, we combine depth completion and metric depth estimation. Specifically, we use DMD3C [17] for depth completion and Video Depth Anything [6] for depth estimation, allowing each model to independently predict depth from the input images. We then apply morphological operations to merge the two predictions: regions covered by LiDAR points are filled using the completion results, while all remaining areas adopt the estimated depth. We do not perform weighted fusion between the two depth maps, as such blending was found to introduce more hallucinations.

Training details of GeoV2V. Our model is built upon Wan2.1-1.3B and trained using approximately 5,000 pairs of synthetic videos, each consisting of 81 frames, with a resolution of 480×832 . Training was conducted on 8 A100 GPUs with a batch size of 1 and gradient accumulation set to 6, for a total of 2,000 steps. The entire training process took approximately 9 days.

Implementation details of FreeVS [37]. In our experiments with FreeVS, we retained the resolution used in the official implementation, as deviating from it significantly degrades performance. We evaluated the model on both the Waymo dataset and the Para4D Synthetic Dataset. To ensure comparable performance on Para4D, we increased the number of merged point cloud frames from 5 to 11, ensuring sufficient point cloud density for reliable processing. However, the LiDAR point clouds in the nuScenes dataset are significantly sparser, making it infeasible to obtain sufficient density through simple frame merging. As a result, FreeVS fails to produce stable or meaningful generations on nuScenes under this setting.

Implementation details of reconstruction-based methods. For StreetGS [44] and OmniRe [7], we follow the official implementation provided by OmniRe [7] and train the models using 81-frame clips under the default configurations. Since EmerNeRF [46] does not support interpolation training on the nuScenes dataset [4], we only evaluate it on the Waymo and Para4D datasets. EmerNeRF is also trained on 81-frame sequences for each experiment.

2 Additional Results

Qualitative Results. Our qualitative results are shown in Figs. 9, 11, 13, and 12. It can be observed that our method outperforms other reconstruction-based approaches in terms of visual quality.

Zero-shot Evaluation on nuScenes. We directly applied our model to perform frame interpolation on the nuScenes dataset. The results are summarized in Table 5, demonstrating that our model exhibits strong generalization capabilities across different datasets and LiDAR feature distributions. Especially in scenarios where the ego vehicle is moving at higher speeds—resulting in sparser sampled viewpoints—our method maintains superior performance, as shown in

Fig. 12. This robustness stems from the fact that our geometry priors are constructed from a single frame at the same (or neighboring) timestamp, making them naturally resilient to large inter-frame viewpoint gaps.

Method	PSNR	SSIM	LPIPS
StreetGS [44]	24.03	0.670	0.336
OmniRe [7]	23.86	0.650	0.301
Ours	23.86	0.658	0.144

Table 5: Zero-shot comparison on the nuScenes dataset for the interpolation task. This evaluation protocol follows the standard setup commonly used to evaluate reconstruction-based methods.

Video Comparison. We provide a video comparison of novel view sequences generated by our method and baseline approaches on a Waymo dynamic driving scenario. Please refer to the accompanying video file submitted as supplementary material for a detailed visual comparison.



Fig. 9: More NVS results of GeoV2V on the Waymo Open dataset. GeoV2V generalizes well across different camera viewpoints and produces consistent novel-view synthesis under varying viewing angles.



Fig. 10: GeoV2V zero-shot on both translation and rotation.



Fig. 11: Qualitative frame-interpolation results on the Waymo Open dataset. While interpolation settings are typically favorable to reconstruction-based methods, our approach still achieves superior visual quality, demonstrating strong performance not only in translation (extrapolation) tasks but also in interpolation scenarios.



Fig. 12: Qualitative frame-interpolation results on the nuScenes dataset.

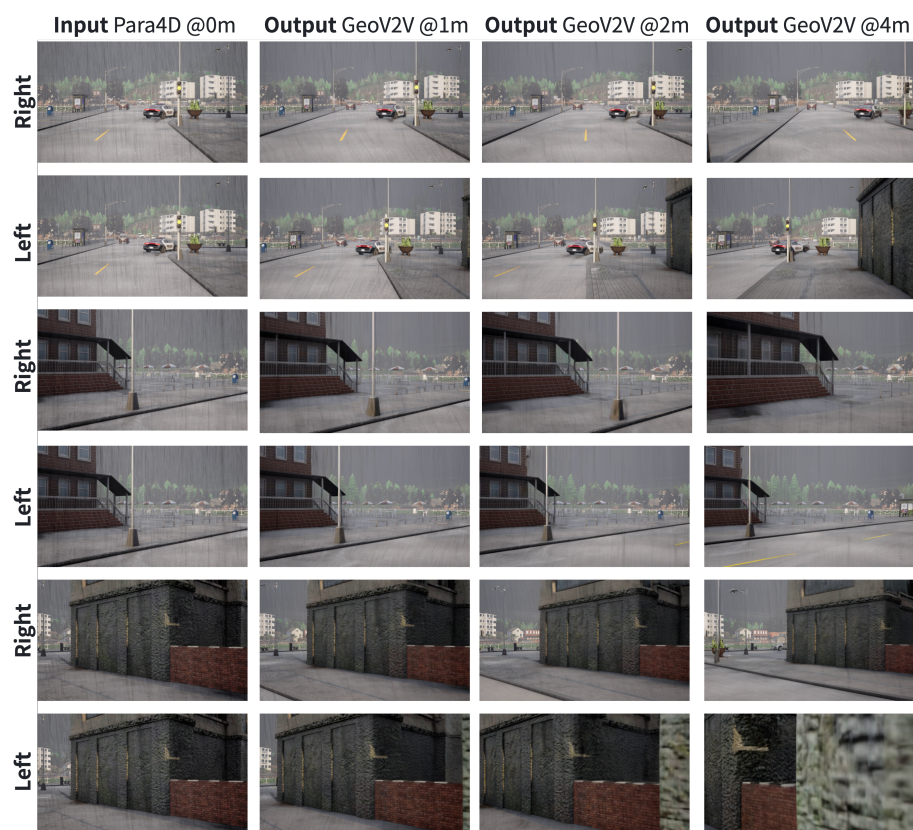


Fig. 13: Qualitative novel view synthesis results of GeoV2V on the Para4D synthetic dataset.

3 Para4D Dataset Details

In this section we provide additional details of the Para4D data organization and camera configuration used in training and evaluation.

Camera configuration. We follow the camera setup used in the Waymo Open Dataset [31]. Each ego vehicle is equipped with five cameras covering the left, front-left, front, front-right, and right views. This configuration provides a wide field of view while maintaining sufficient overlap between adjacent cameras, which facilitates stable cross-view correspondence and view synthesis. All cameras have the same intrinsic parameters as those in the Waymo dataset and are rigidly mounted on the ego vehicle, which ensures consistent multi-view observations over time.

Multi-view temporal observations. Fig. 15 illustrates the multi-view observations recorded in Para4D. For each trajectory, synchronized images are captured simultaneously from the five cameras at every timestamp. Each column corresponds to a camera view, while rows represent consecutive temporal frames from different dynamic driving scenes. This multi-view and temporally aligned structure provides consistent observations of moving vehicles, traffic signals, and environmental changes, which is essential for learning temporally coherent view synthesis models.

Parallel-lane observations. Together with standard multi-view capture, Para4D provides synchronized observations from multiple laterally shifted trajectories. As shown in Fig. 16, cameras are placed on parallel lanes with lateral offsets ranging from -4 m to $+4$ m relative to a reference trajectory. Each column represents a different lane offset, while rows correspond to frames from the same driving sequence.

This design provides strong pixel-level geometry-guided supervision for video-to-video learning across neighboring lanes while preserving the same dynamic scene evolution. In contrast, previous synthetic datasets such as XLD [16] capture fewer camera views and mainly focus on static environments, typically consisting of empty maps without dynamic agents such as vehicles or pedestrians, shown in Fig. 14. These synchronized cross-lane observations enable video-to-video models to learn consistent viewpoint transitions under complex 4D dynamic scenes.



Fig. 14: Lane-offset views from previous synthetic dataset XLD [16]. While providing laterally shifted viewpoints, the scenes remain static without dynamic objects or temporal evolution.



Fig.15: Multi-view camera observations in the Para4D dataset. Para4D simultaneously captures synchronized images from five ego-vehicle camera views for each parallel trajectory, following the same camera configuration as the Waymo dataset. Each row shows consecutive temporal frames from different driving scenes. This multi-view and temporally consistent setup is designed to support the training and evaluation of NVS methods.

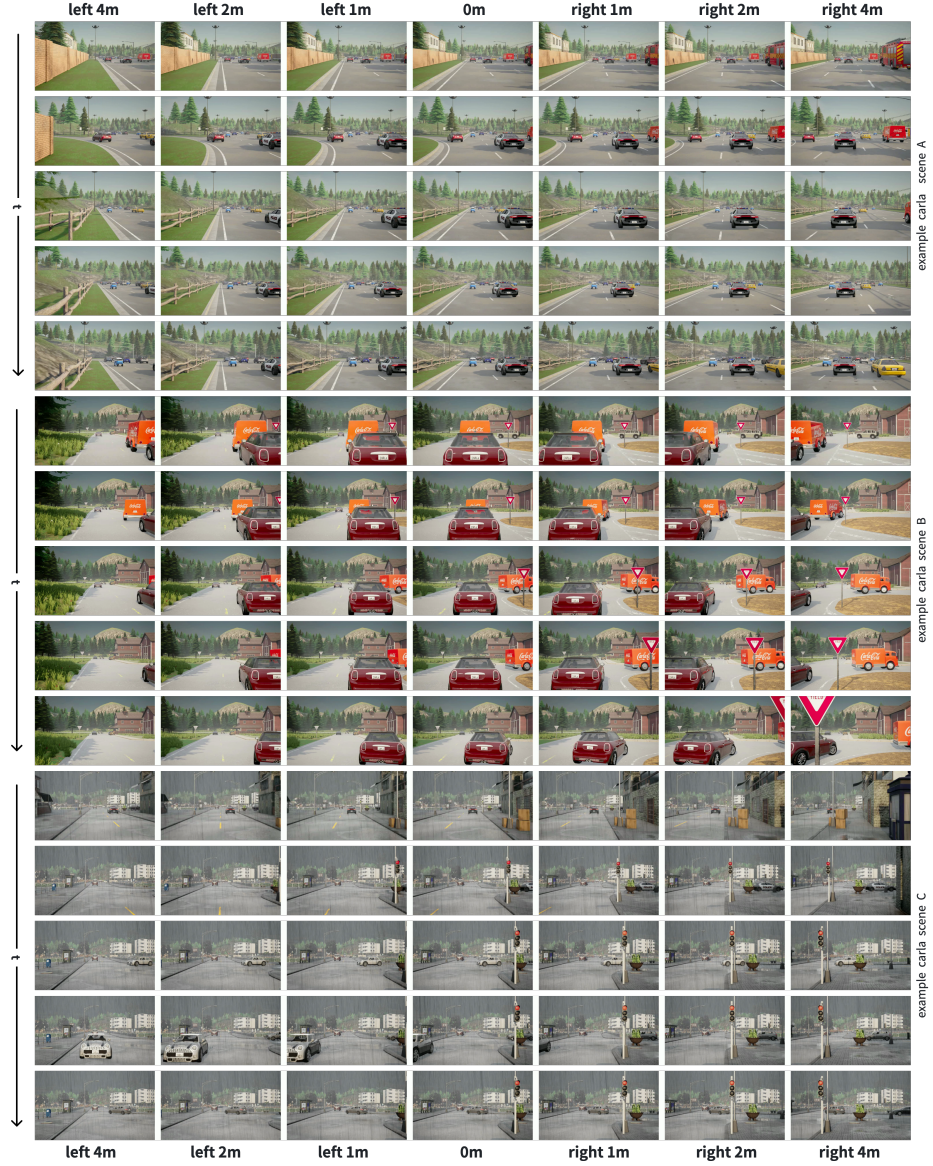


Fig. 16: Pixel-Level Geometry-Guided V2V Training supported by Para4D. Para4D makes cross-lane video-to-video learning feasible in complex 4D dynamic driving scenes by providing synchronized multi-lane observations and strong pixel-level geometric supervision. Ablation results show that without this supervision, novel-view video generation loses reliable geometric control, leading to synthesized views that are not consistently aligned with the original viewpoint and thus produce lower PSNR.



Fig. 17: Visualization of the domain gap limitation on real-world data.



Fig. 18: Visualization of the temporal inconsistency caused by WanVAE’s temporal cache mechanism. The image quality of the $4n$ -th frame is significantly better than that of the $(4n + 1)$ -th frame, exemplified by the clarity of the “CHASE” text, which is severely blurred in the $(4n + 1)$ -th frame.

4 Limitations

During our research, we identified several limitations of our proposed approach.

First, since our end-to-end video diffusion model is primarily trained on large-scale synthetic data, there still exists a certain domain gap when applied to real-world datasets. At a distance of around 4 m, heavy occlusion causes the model to lack sufficient context for inferring the actual color of the scene. As shown in Fig. 17, the model may produce regions with plausible textures but incorrect colors.

Second, the WanVAE we adopted employs a temporal compression scheme, which introduces certain motion blur artifacts when encoding scenes with high dynamics and relatively low frame rates, typical of autonomous driving scenarios. This is mainly reflected in the fact that every 4th frame has noticeably better quality than the other frames, as shown in Fig. 18, especially under large optical flow.

Third, similar to ReCamMaster [2], our method adopts a frame-wise concatenation strategy, which leads to the trained model being sensitive to the input

sequence length. Consequently, the number of frames during inference must be strictly consistent with that used during training.

Our comparison is system-level rather than strictly backbone-controlled, as GeoV2V and several baselines use different generative backbones. However, ablations on Para4D training, geometric priors, and full-video conditioning show that the gains do not come from the backbone alone.

5 Experiment Details

Scenes for Validation. For the nuScenes dataset, we select 10 scenes from the v1.0-mini split as our validation set, matching the standard mini-dataset configuration. The scene IDs are: scene-0061, scene-0103, scene-0553, scene-0655, scene-0757, scene-0796, scene-0916, scene-1077, scene-1094, and scene-1100.

For the Waymo dataset, we generally follow the static/dynamic split defined in NOTR [46]. Since our evaluation requires conducting NVS on dynamic scenarios, we adopt the same 12 spacious sequences selected by FreeVS [37] as our dynamic validation set. These sequences provide sufficient camera motion and scene complexity for reliable assessment. The selected validation sequences from the WOD dataset are:

```
segment-10588771936253546636_2300_000_2320_000,
segment-6242822583398487496_73_000_93_000,
segment-16801666784196221098_2480_000_2500_000,
segment-1191788760630624072_3880_000_3900_000,
segment-10625026498155904401_200_000_220_000,
segment-11846396154240966170_3540_000_3560_000,
segment-18111897798871103675_320_000_340_000,
segment-11017034898130016754_697_830_717_830,
segment-10963653239323173269_1924_000_1944_000,
segment-12161824480686739258_1813_380_1833_380,
segment-11928449532664718059_1200_000_1220_000,
segment-10444454289801298640_4360_000_4380_000.
```

6 Supplementary Videos

The supplementary materials include several videos that demonstrate the qualitative performance of our method as well as the characteristics of the Para4D dataset. Below we briefly describe the content of each video.

- **GeoV2V_Main_Comparison.mp4** This video presents a comparison between our method and representative reconstruction-based and reconstruct-then-restore approaches. The comparison is conducted on real-world sequences from the Waymo dataset.
- **GeoV2V_Main_Results.mp4** This video showcases the view translation results produced by our model on real-world Waymo data.

- **GeoV2V_Tested_On_Para4D.mp4** This video presents the performance of our model when applied to the synthetic Para4D dataset. It demonstrates the effectiveness of the proposed method under controlled conditions and different view translation offsets.
- **Para4D_Dataset_Visualization_Simple.mp4** This video highlights the main characteristics of the Para4D dataset, including parallel viewpoints and dynamic scenes. It provides a concise overview of the dataset’s core properties.
- **Para4D_Dataset_Visualization_Full_Cameras.mp4** This video visualizes the Para4D dataset captured with multiple parallel camera rigs. It provides an overview of the synchronized multi-camera setup and illustrates the diversity of captured scenes.