

RecEdit-Drive:3D Reconstruction-Guided Spatiotemporal Video Editing for Autonomous Driving Scenes

Supplementary Material

1. Spatial Position Control

Table 1. The spatial control capability of RecEdit-Drive is evaluated using 3D object detection metrics, while the semantic consistency between the edited foregrounds and the reference images is measured with CLIP-I.

Tasks	mRecall \uparrow	mATE \downarrow	mAOE \downarrow	CLIP-I \uparrow
Replacement	0.971	0.50	0.041	77.53
Insertion	0.979	0.44	0.040	77.61
Repositioning	0.967	0.48	0.042	77.51
Original	0.986	0.43	0.038	77.64

To evaluate the spatial control capability of RecEdit-Drive in foreground editing, we perform 3D object detection on the edited data using StreamPETR. Since the target object is removed in the deletion task, detection is conducted only on the data edited by the replacement, insertion, and repositioning tasks. As shown in Table 1, the edited results reveal only a slight performance drop compared to the unedited original data, demonstrating that RecEdit-Drive maintains robust spatial position control and semantic alignment for the edited foregrounds across all editing tasks.

2. Computational Cost Analysis

Table 2. Comparison of Computational Cost between RecEdit-Drive and DriveEditor.

Method	GPU Mem	Time	Param
ProPainter	10.96GB	0.515s/frame	39.4M
SD	9.90GB	11.155s/frame	440.5M
T2V	8.44GB	1.988s/frame	1.066B
TAV	4.99GB	147.124s/frame	1.116B
DriveEditor	30.97GB	8.95s/frame	4.2B
RecEdit-Drive	33.93GB	9.99s/frame	4.5B

We use an A6000 to generate videos and report computational costs in Table 2. SV3D and homography warping account for about 18% and 8% of the total runtime, respectively. In RecEdit-Drive, homography-based feature construction is purely geometric and parameter-free. Additional costs mainly arise from attention operations in our SFW module.

We can generate diverse realistic videos that are difficult to collect in real-world settings (e.g., wrong-way driving and dangerous lane changes). These scenes are critical for tasks such as object detection and tracking. Considering inference time (~ 10 s/frame), RecEdit-Drive requires only 11 hours to create a large-scale autonomous driving dataset comparable to nuScenes using 8 A6000 GPUs, highlighting its strong practical utility.

3. Ablation of RecEdit-Drive

3.1. Ablation of Soft Attention Mask

Table 3. Ablation study on Soft Attention Mask.

Methods	FID \downarrow	FVD \downarrow
w/o soft attention mask	5.58	16.37
w soft attention mask	5.22	14.38

In the SCM module, the soft attention mask $\mathbf{M}^{\mathbf{F},\mathbf{G}}$ is used to compute the attention guidance mask \mathcal{M} , which facilitates the natural blending of the edited foreground with the background. In Table 3, we evaluate the variant that removes the Gaussian kernel smoothing and directly computes $\hat{\mathcal{M}} \in \mathbb{R}^{HW \times HW}$ from $\mathbf{M}^{\mathbf{F}}$ as follows:

$$\hat{\mathcal{M}}_{i,j} = C(1 - \mathbf{M}_i^{\mathbf{F}} \odot \mathbf{M}_j^{\mathbf{F}}), (1 \leq i, j \leq N), \quad (1)$$

where C is a negative constant that satisfies $C \ll 0$. i and j are the indices of video frames. As shown in Table 1, removing the soft mask leads to a significant degradation in both FID and FVD. This is because $\hat{\mathcal{M}}$ forces the model to focus its attention solely on the edited region, resulting in abrupt attention transitions at the boundaries between edited and non-edited areas. These discontinuities produce unnatural foreground-background blending, thereby degrading the image quality of individual frames and the temporal consistency across video sequences.

3.2. Number of Reference Views

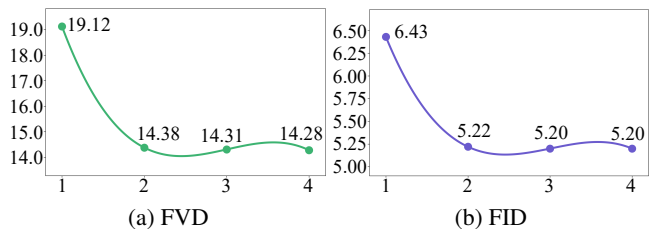


Figure 1. Comparison of the effects of using different numbers of reference views in the Spatial Feature Warping (SFW) module to construct 3D prior information on model performance. FVD and FID results are reported on the evaluation set.

As shown in Figure 1, the model performance drops significantly in the SFW module when only the nearest novel view is used as the reference view, yielding an FID of 19.12 and an FVD of 6.43. This degradation occurs because a single reference view is insufficient to construct a complete 3D prior, making it difficult for the model to generate accurate foreground edits. In contrast, using two reference views already provides all the essential information required for the

target viewpoint, and consequently, increasing the number of reference views further does not lead to noticeable performance improvements.

3.3. Different Timestep Thresholds on Noise Replace

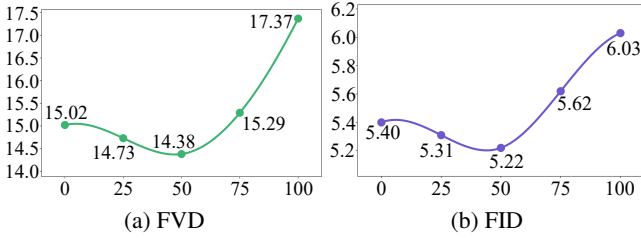


Figure 2. Comparison of the effects of the timestep threshold used in the Noise Replace mechanism on model performance. The x-axis denotes the percentage of early denoising steps during which background noise is replaced. FVD and FID metrics are reported on the evaluation set.

As shown in Figure 2, the Noise Replace module achieves the best performance when background noise from the corresponding forward-noising timestep is used to replace the background of the denoised output during the first 50% ($t > T/2$) of the denoising process, yielding an FVD of 14.38 and an FID of 5.22. As the timestep threshold increases, the model performance gradually deteriorates. Although performing background-noise replacement more frequently enforces stronger consistency in non-edited regions, excessive replacement introduces unnatural boundaries between the edited foreground and the background, which severely degrades visual quality and spatiotemporal coherence. In contrast, reducing the timestep threshold also leads to some performance drop, but the impact is relatively minor. Since the primary purpose of background-noise replacement is to preserve non-edited regions, fewer replacements may cause slight deviations in the background but have limited effect on the visual fidelity and temporal consistency of the edited results.

3.4. Number of Cross-Frame Inputs

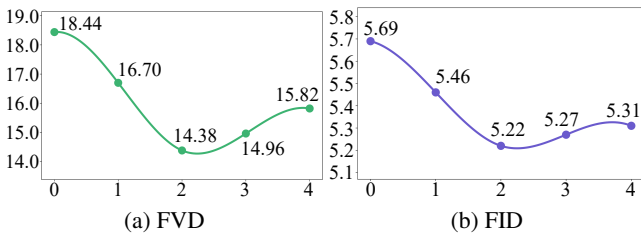


Figure 3. Effects of modeling spatiotemporal correlations with varying numbers of neighboring frames in the Spatiotemporal Collaborative Modeling (SCM) module. FVD and FID results are evaluated on the validation set.

In the SCM module, we evaluate the effects of modeling spatiotemporal correlations between the cur-

rent frame z_n and the adjacent-frame sequence $Z = \{z_{n-1}, z_{n+1}, z_{n-2}, z_{n+2}, \dots\}$, where the first N frames in Z are selected for cross-frame attention. As shown in Figure 3, the model achieves the best performance when $N = 2$. When N is too small, the limited temporal context is insufficient for reliable propagation of scene information, making it difficult to maintain spatiotemporal consistency in the edited results. In contrast, when N becomes too large, the visual discrepancies between distant frames introduce inconsistent temporal information, which can degrade the quality of the edited outputs.

3.5. Synthetic Data Scaling and Saturation

As shown in Figure 4, we assess performances of downstream 3D object detection using various proportions of synthetic data. The performance increases steadily as the proportion of synthetic data rises from 0% to 100%, while the improvements become marginal as more synthetic data is added. The performance gains mainly arise because the edited data provide richer and geometrically consistent vehicle pose variations and enable the creation of key interaction scenarios (e.g., tightly packed vehicles) that benefit detection performance on the closed test set. As vehicle pose distribution is increasingly covered, the marginal benefit of additional data diminishes, gradually saturating performance. Notably, in more challenging detection scenarios involving open and risky interactions, our method can synthesize such cases to further improve detection performance.

4. 3D Box and 2D Mask Acquisition

3D box and 2D mask sequences are both used during inference. 3D boxes are derived from user inputs, which specify the category and intended trajectory of the edited object. We develop a tool, as shown in Figure 5, to map these inputs on the BEV map to 3D boxes, indicating object location and velocity direction. If obtained 3D boxes overlap with existing foreground objects, the occluded objects will be removed to preserve overall scene realism. 2D masks are obtained via SAM to indicate the target deletion regions. They do not require precise alignment with object boundaries. Given inaccurate masks (e.g., uncovered region $\leq 10\%$, as shown in Figure 6) our SCM allows the model to leverage temporal context across frames, resulting in coherent and visually consistent outputs.

5. Failure case

Due to the interactive nature of our method, performance can be affected by the quality of input 2D masks. We conduct a quantitative analysis of cases, as shown in Figure 6, in which the mask fails to fully cover the target foreground object in some frames, for assessing the robustness to mask

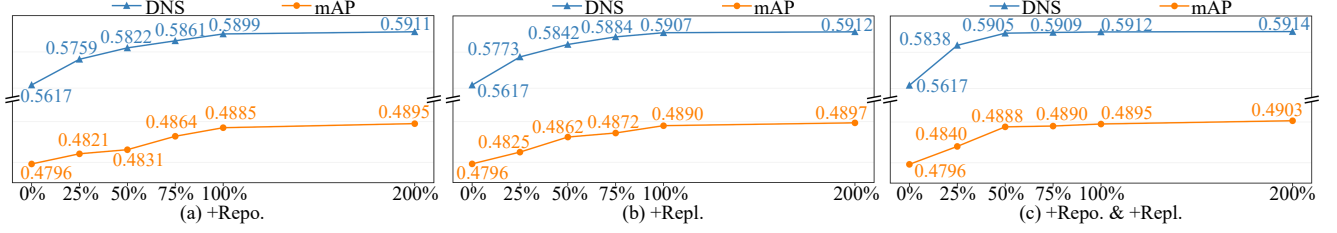


Figure 4. Synthesis data ratio for downstream task performance. (a) Adding only Repo. data; (b) adding only Repl. data; (c) adding equal amounts of Repo. and Repl. data. The x-axis denotes the ratio of synthetic data to real data.



Figure 5. 3D boxes and 2D masks acquisition tool. (a) In the BEV view, the red arrow originates from the center of the 3D bounding box and indicates its orientation. The user-annotated trajectory is accumulated in the BEV view of the next frame. (b) The blue box indicates the foreground object being edited.

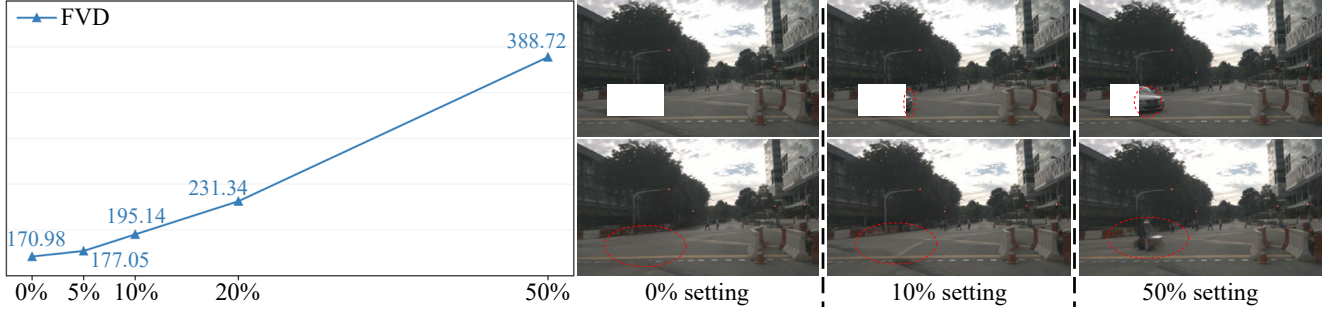


Figure 6. Robustness Analysis of Performance to Mask Accuracy. The x-axis denotes the percentage of the target foreground area uncovered by the mask in a single frame of the mask sequence.

deviations and extending the analysis of failure cases. Our SCM ensures the generation of visually consistent content even when the uncovered region is noticeable (e.g., uncovered region of the fg object is 10%), while the performance degrades as the uncovered region increases (>10%, which is rare for SAM). A promising direction is to leverage semantic representations extracted from the input masked video as additional guidance to correct erroneous video features.

6. Additional Visual Comparison

In Figures 7, 8, 9, and 10, we provide additional visual comparisons between our RecEdit-Drive and other methods. The comparisons show the effectiveness of our RecEdit-Drive, which surpasses DriveEditor across deletion, replacement, insertion, and reposition tasks.

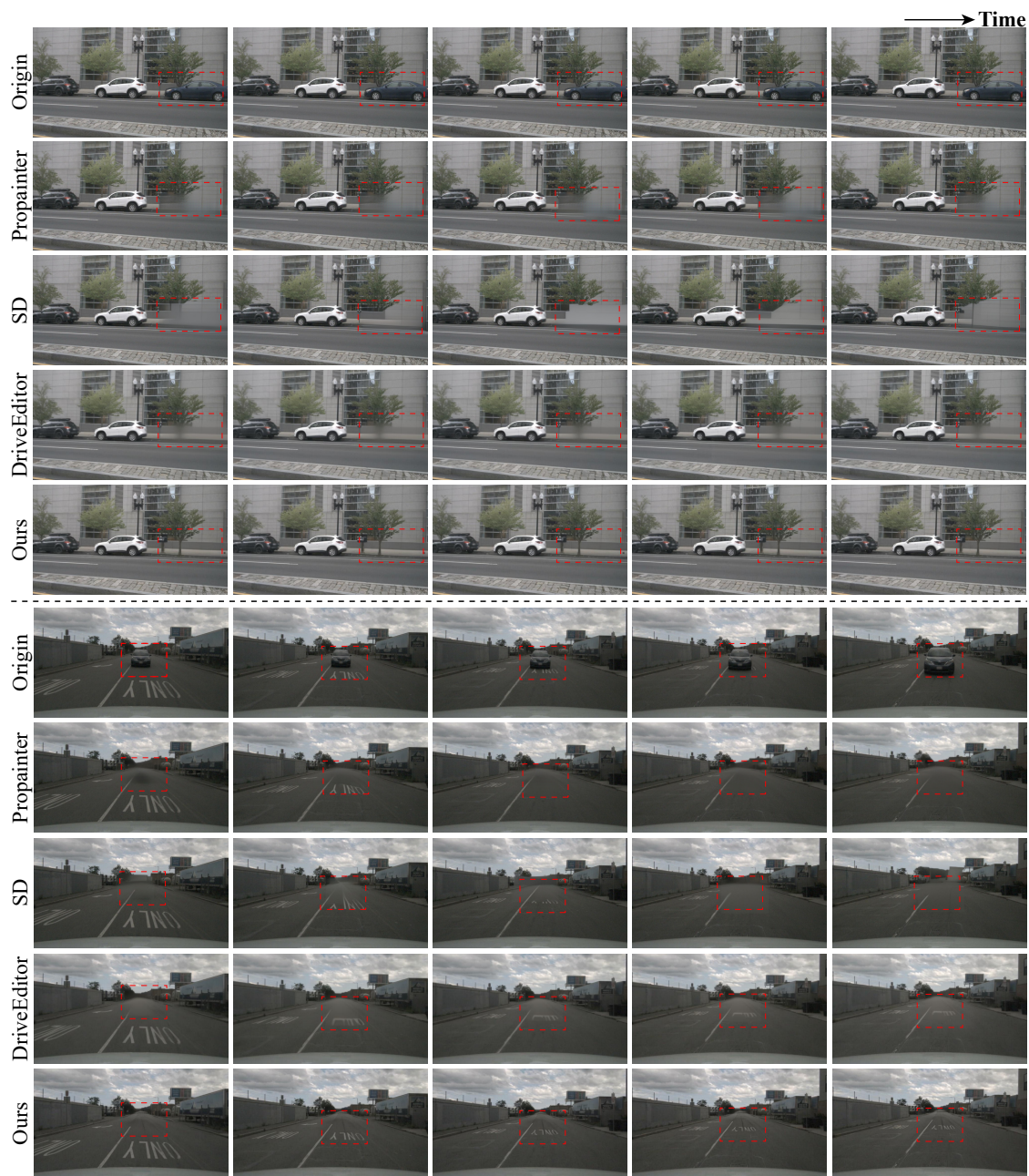


Figure 7. Comparison of deletion results between DriveEditor and other methods.

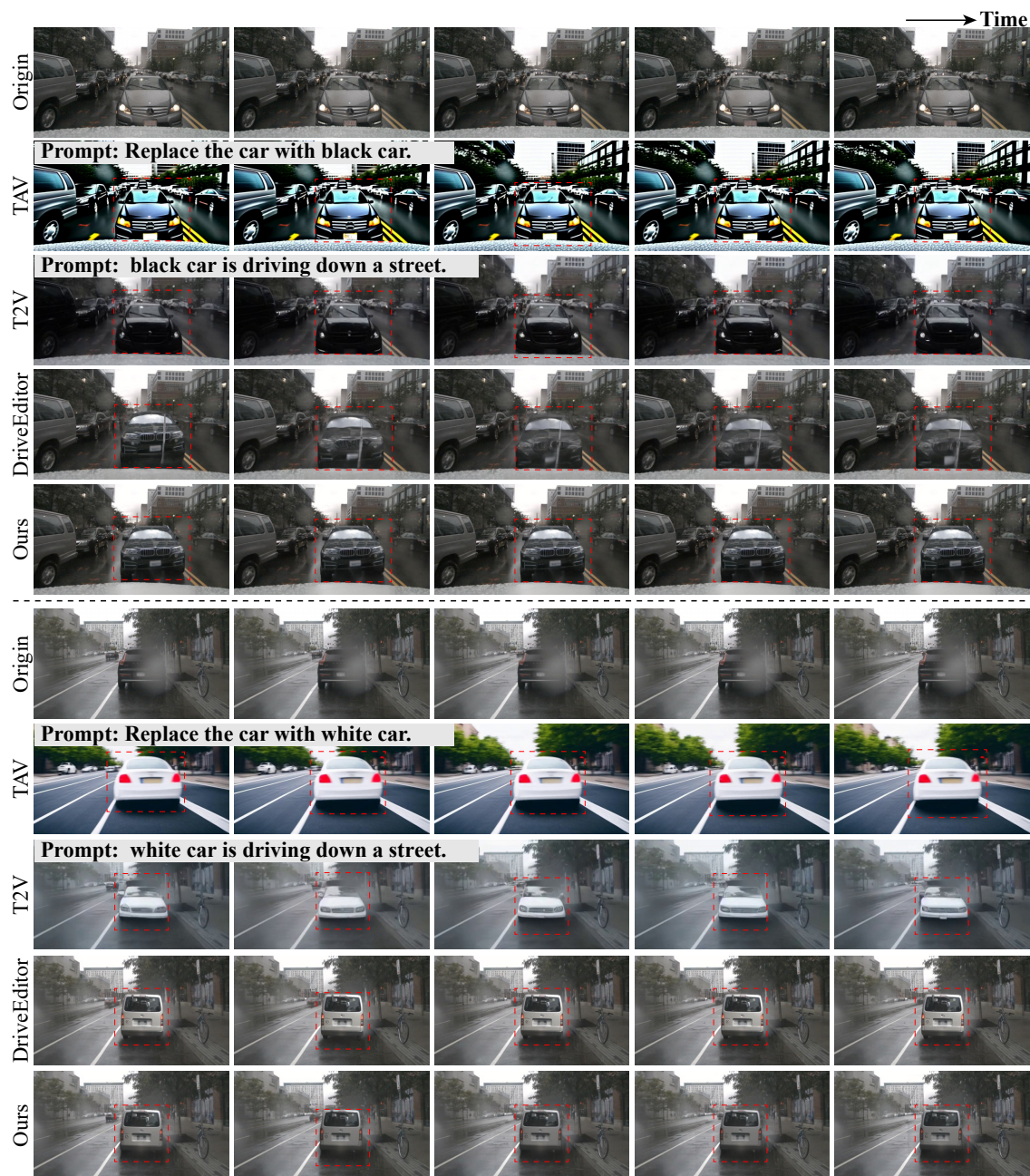


Figure 8. Comparison of replacement results between DriveEditor and other methods.



Figure 9. Comparison of repositioning results between DriveEditor and RecEdit-Drive.

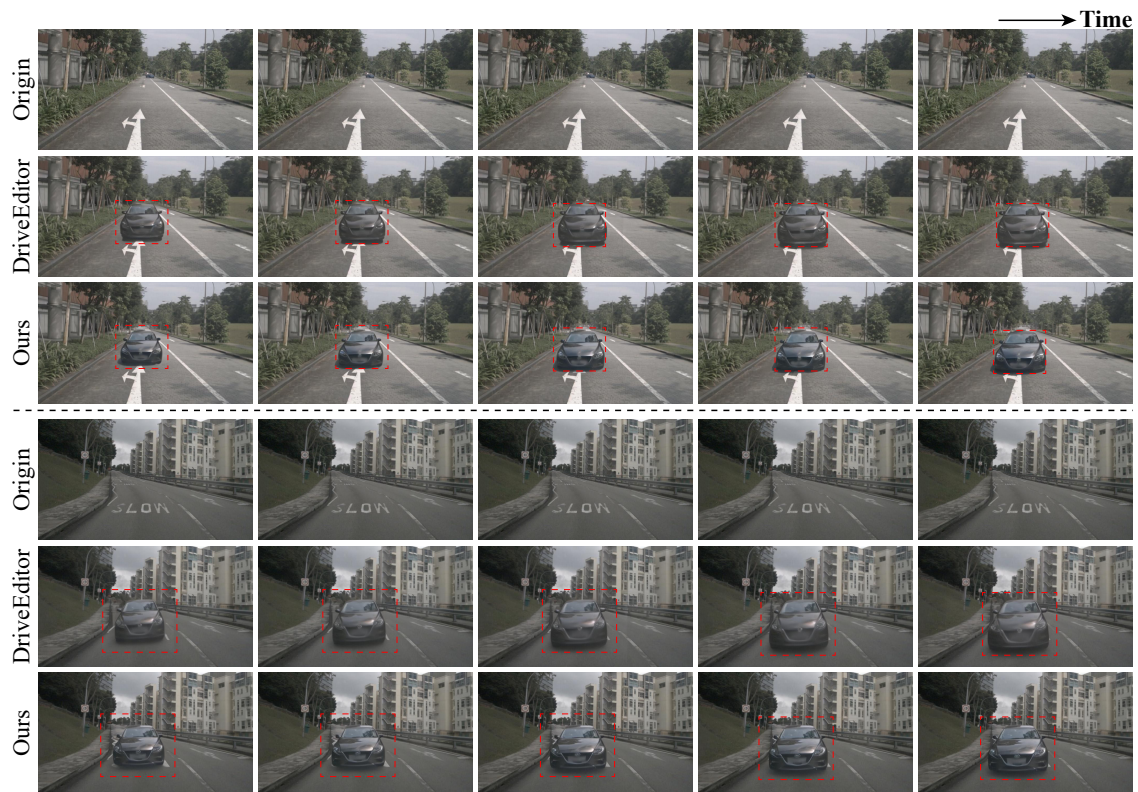


Figure 10. Comparison of insertion results between DriveEditor and RecEdit-Drive.