MoVieDrive: Urban Scene Synthesis with Multi-Modal Multi-View Video Diffusion Transformer

Supplementary Material

In this supplementary material, we discuss the limitation of this work in Section 6, present more discussions in Section 7, and present additional visualizations of multi-view multi-modal generation results in Section 8.

6. Limitation and Future Work

While our approach has achieved superior multi-modal multi-view generation performance, there are still some limitations. First, how to effectively combine the proposed approach with a closed-loop autonomous driving simulator is worth further study. This potentially helps to comprehensively assess the safety and reliability of autonomous driving systems. Second, how to derive LiDAR point clouds from the generated multi-modal data is an interesting research direction to further enhance our approach. Moreover, as discussed in Section 4.5, although our approach can be extended for long video generation, there is still room for improvement in terms of long video quality. Our future work aims to address these problems to facilitate the deployment of our approach for real-world applications.

7. More Discussions

Is this a straightforward modification of CogVideoX with cross-view attention? Our approach is not a straightforward modification of CogVideoX [40] for driving scene video generation. Although we use temporal layers and 3D VAE from CogVideoX, we significantly modify them with a unified multi-modal multi-view diffusion transformer model for autonomous driving scene generation and propose efficient diverse conditioning inputs encoding to improve scene video generation quality and controllability. A straightforward modification of CogVideoX is adding cross-view attention layers, specific heads, and our conditions to CogVideoX. However, as shown in Tab. 5, this yields worse performance compared with our design. Another simple variant is using the fixed CogVideoX and 3D VAE with learnable modal-specific heads but using neither our multi-view spatiotemporal blocks nor cross-view attention layers. From Tab. 5, we can see that this simple variant achieves significantly worse performance. In addition, combining CogVideoX with SyntheOcc [15] still yields worse results compared with our approach.

Is this a simple extension of existing multi-modal synthesis methods? Our approach is not a simple extension of existing multi-modal synthesis methods. Compared with existing methods, the innovation of this work lies in that we propose the first work that exploits diverse conditioning inputs and a unified multi-modal multi-view diffusion transformer model for multi-modal multi-view autonomous driving scene generation. This fills a gap left by existing works, facilitating urban scene understanding in autonomous driving. As discussed in Section 2, none of the existing diffusion-based multi-modal synthesis methods are specifically designed for multi-modal multi-view autonomous driving scene generation. It is not a trivial task to modify these methods for driving scene generation. For example, HyperHuman [18] is one of the recent works that explores diffusion-based multi-modal synthesis, but it is specifically devised for human generation so adapting this method to our problem requires modifying conditioning input encoders, diffusion transformer layers, training processes, *etc*. Similar situations applied to other existing methods. To alleviate this concern, we implement a model variant that employs the modality-specific heads [37], multi-view spatiotemporal blocks, CogVideoX, 3D VAE, and ours conditions. As shown in Tab. 6, this variant still yields worse performance compared with our approach. Although more careful modifications of existing multi-modal synthesis methods may improve their performance, from this result we know that adapting existing general methods to multi-modal multi-view driving scene video generation is not a trivial task. We leave more comprehensive studies for future work.

Can this approach be used with other off-the-shelf models for multi-modal data preparation? Our approach is orthogonal to the used off-the-shelf models. In practice, when groundtruth multi-modal data are provided, our approach can be used to train a unified model for multi-modal multi-view scene generation. Since no groundtruth multi-modal data are provided on the nuScenes and Waymo

Table 5. Comparison with other model vairants on nuScenes.

Methods	FVD↓
Ours	46.8
CogVideoX + Cross-view attention + Our conditions	118.4
Fixed CogVideoX + Learnable head	364.1
CogVideoX + SyntheOcc	60.4

Table 6. Comparison with a modal-specific variant on nuScenes.

Methods	FVD↓
Ours	46.8
Modality-specific head based variant	109.4

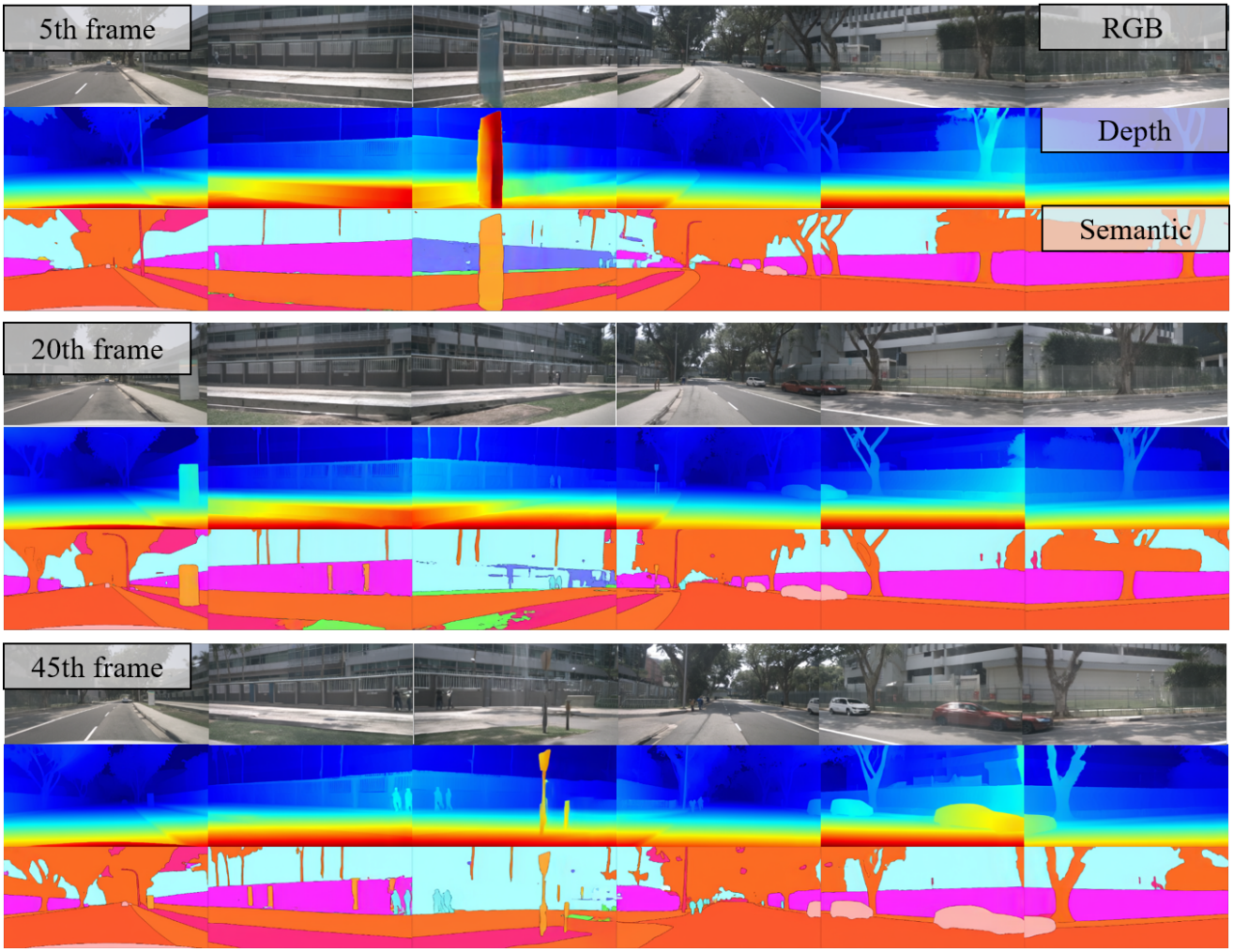

Figure 10. Multi-modal multi-view generation with DVIS++ and VDA for multi-modal data preparation.

datasets, we use Depth-Anything-V2 and Mask2Former in this work. These models are widely used for depth and semantic prediction in autonomous driving scene understanding. It is worth noting that RGB video generation is key for this task. If the quality of generated RGB frames is poor, multi-modal prediction can hardly yield good performance. Thus, the collaboration of multi-modal multi-view generation is a highlight of our approach and Tab. 2 verifies that our unified approach can reduce the number of models used and achieve better overall performance for multi-modal multi-view urban scene synthesis. To alleviate this concern, we further conduct experiments with DVIS++ [42] for semantic maps preparation and VDA [4] for depth maps preparation. Experimental results show that our approach with DVIS++ and VDA yields FVD of 47.4, which is similar to the performance of our approach with Depth-Anything-V2 and Mask2Former (FVD of 46.8). In Fig. 10, we show some results of our approach with DVIS++ and VDA. We can observe that our approach can still generate high-quality multi-modal multi-view urban scene videos with cross-view consistency and cross-modal consistency. Note that colors of the multi-modal maps are depended on the predefined color palette of the off-the-shelf models and can be mapped to calculate the metrics.

8. More Visualization Results

In Figs. 11 to 19, we show additional visualizations of our multi-view multi-modal generation results, including enlarged versions of some figures presented in the main paper.

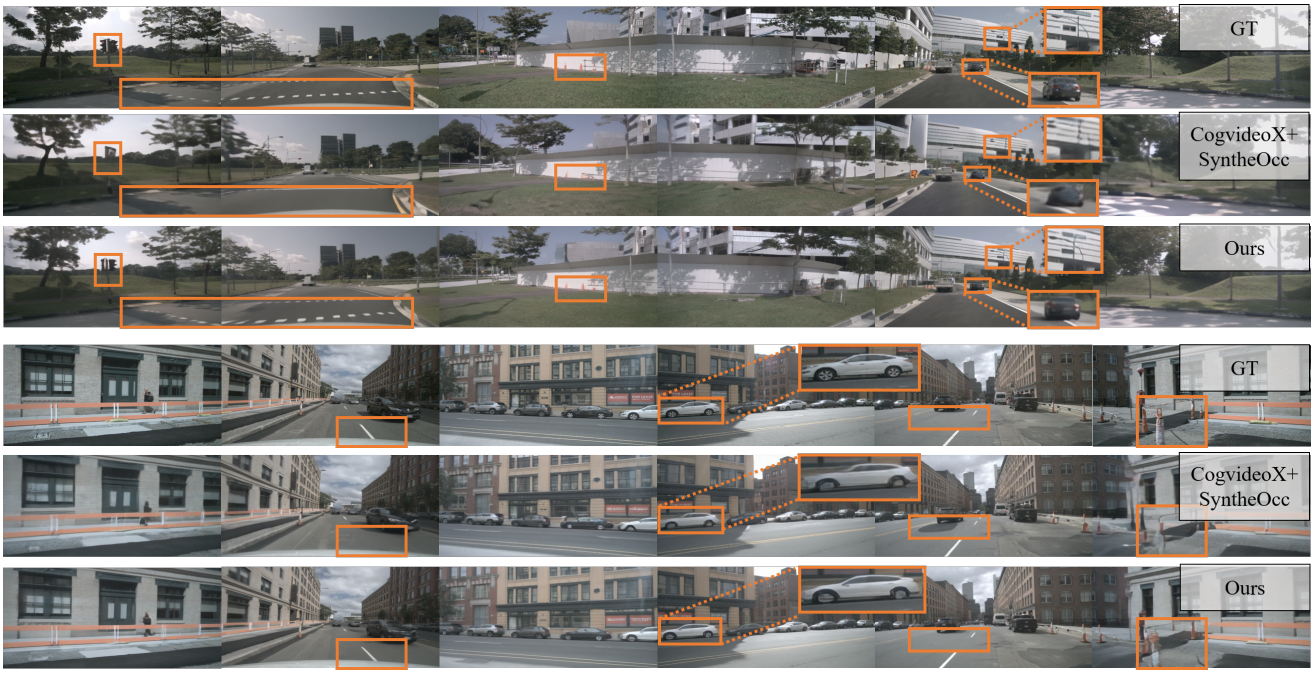

Figure 11. Quantitative comparison on nuScenes (an enlarged version of Fig. 4).

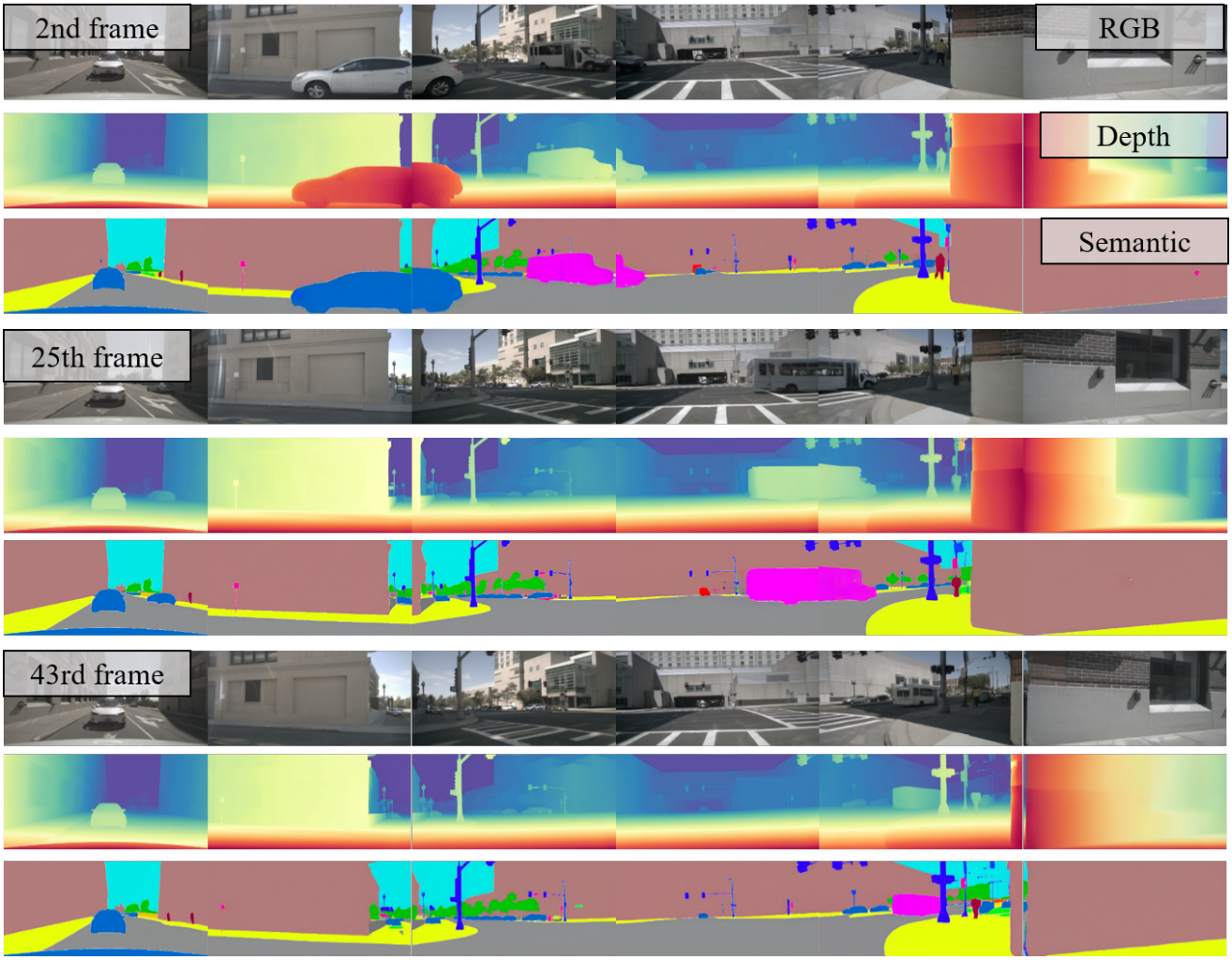

Figure 12. Additional visualizations of our multi-view multi-modal generation results on nuScenes (arbitrarily selected frames covering different time).

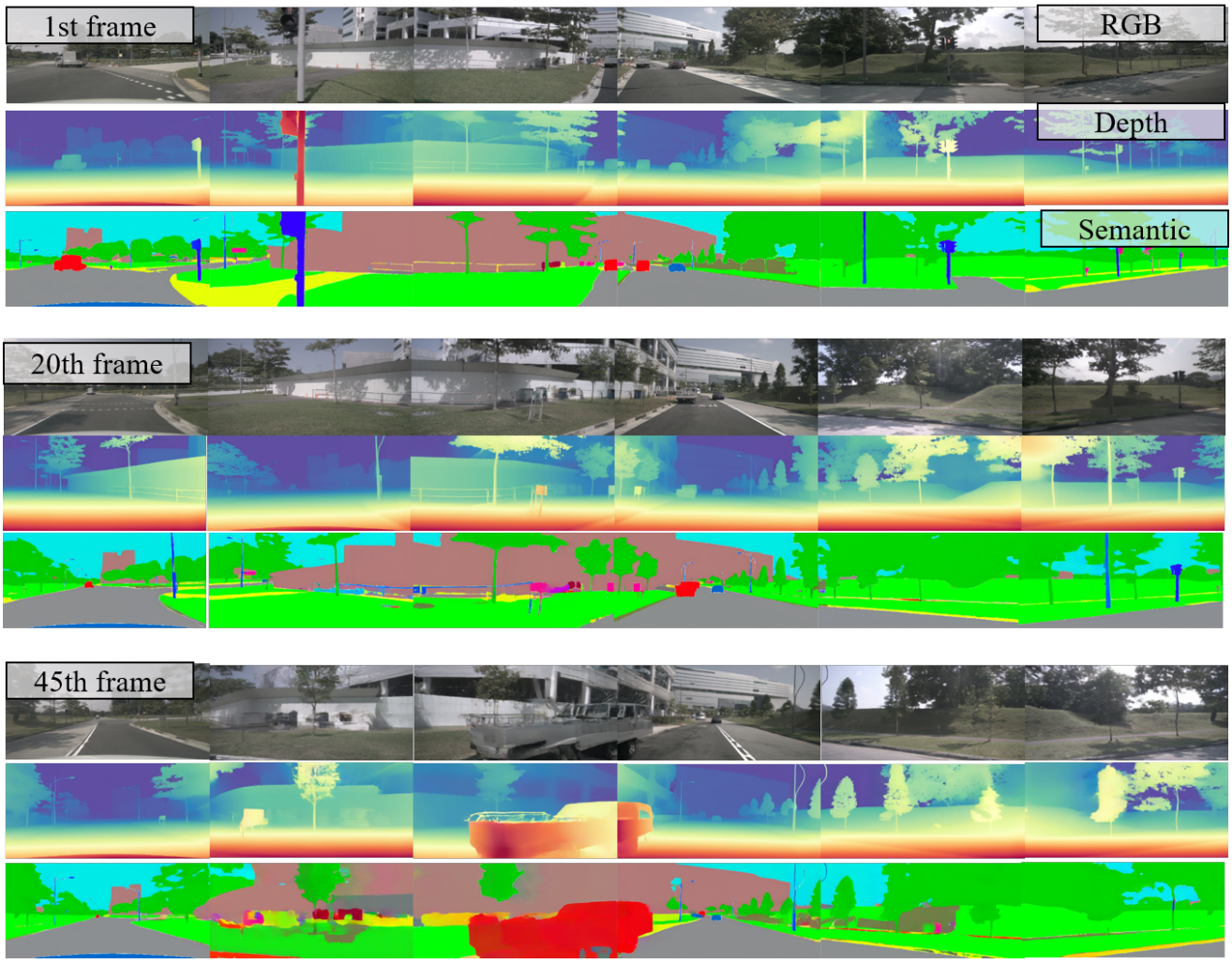

Figure 13. Additional visualizations of our multi-view multi-modal generation results on nuScenes (arbitrarily selected frames covering different time).

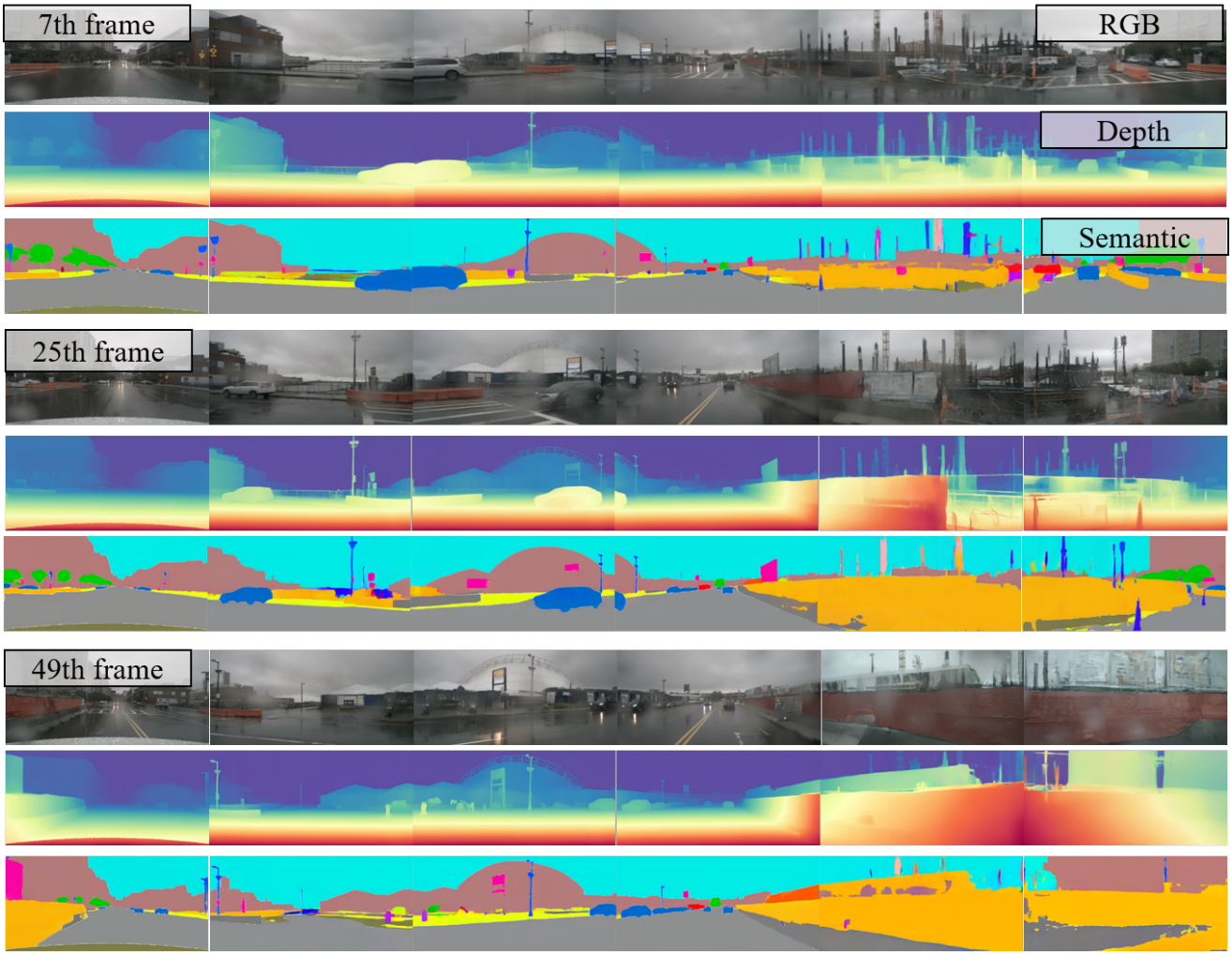

Figure 14. Additional visualizations of our multi-view multi-modal generation results on nuScenes (arbitrarily selected frames covering different time).

Figure 15. Additional driving scene generation with diverse weather conditions on nuScenes.

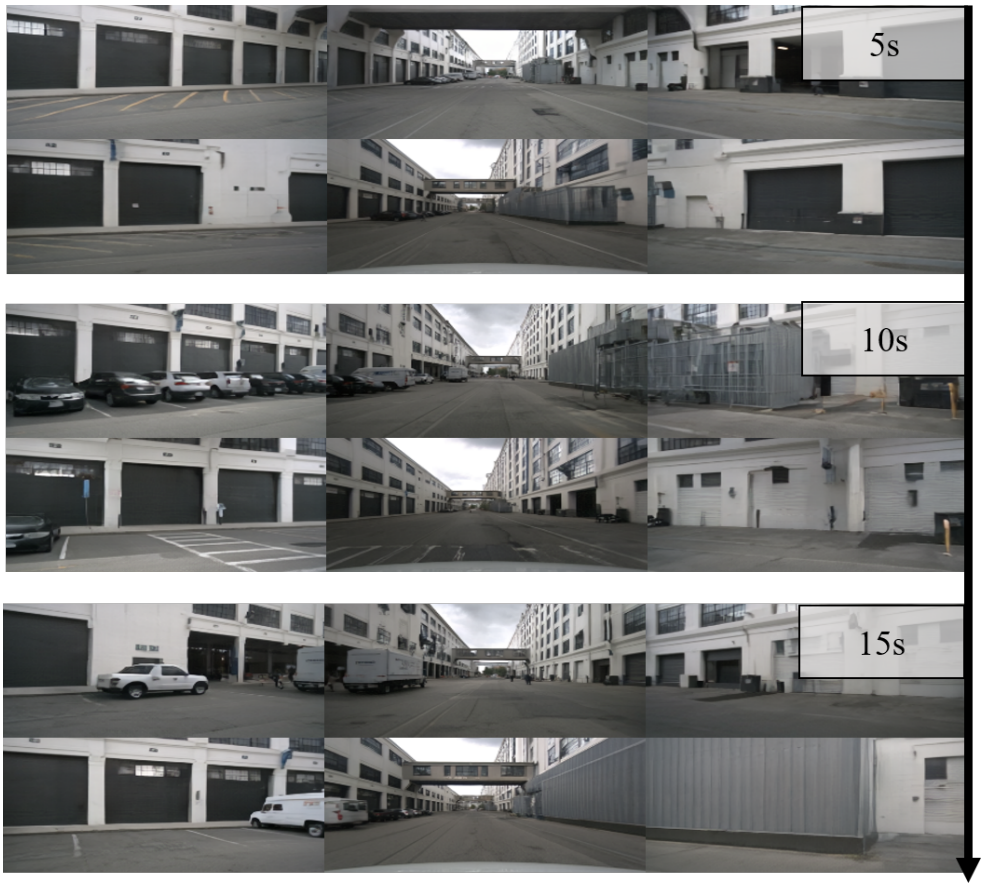

Figure 16. Additional long video generation without reference frames on nuScenes.

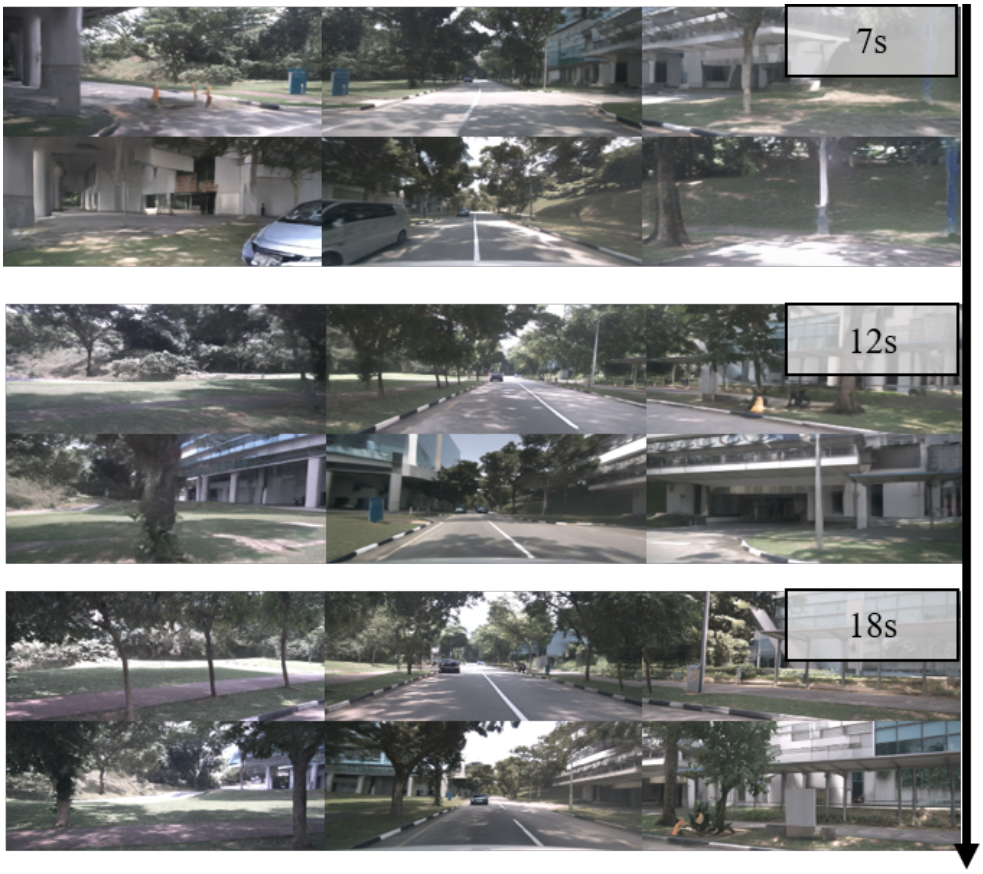

Figure 17. Additional long video generation without reference frames on nuScenes.

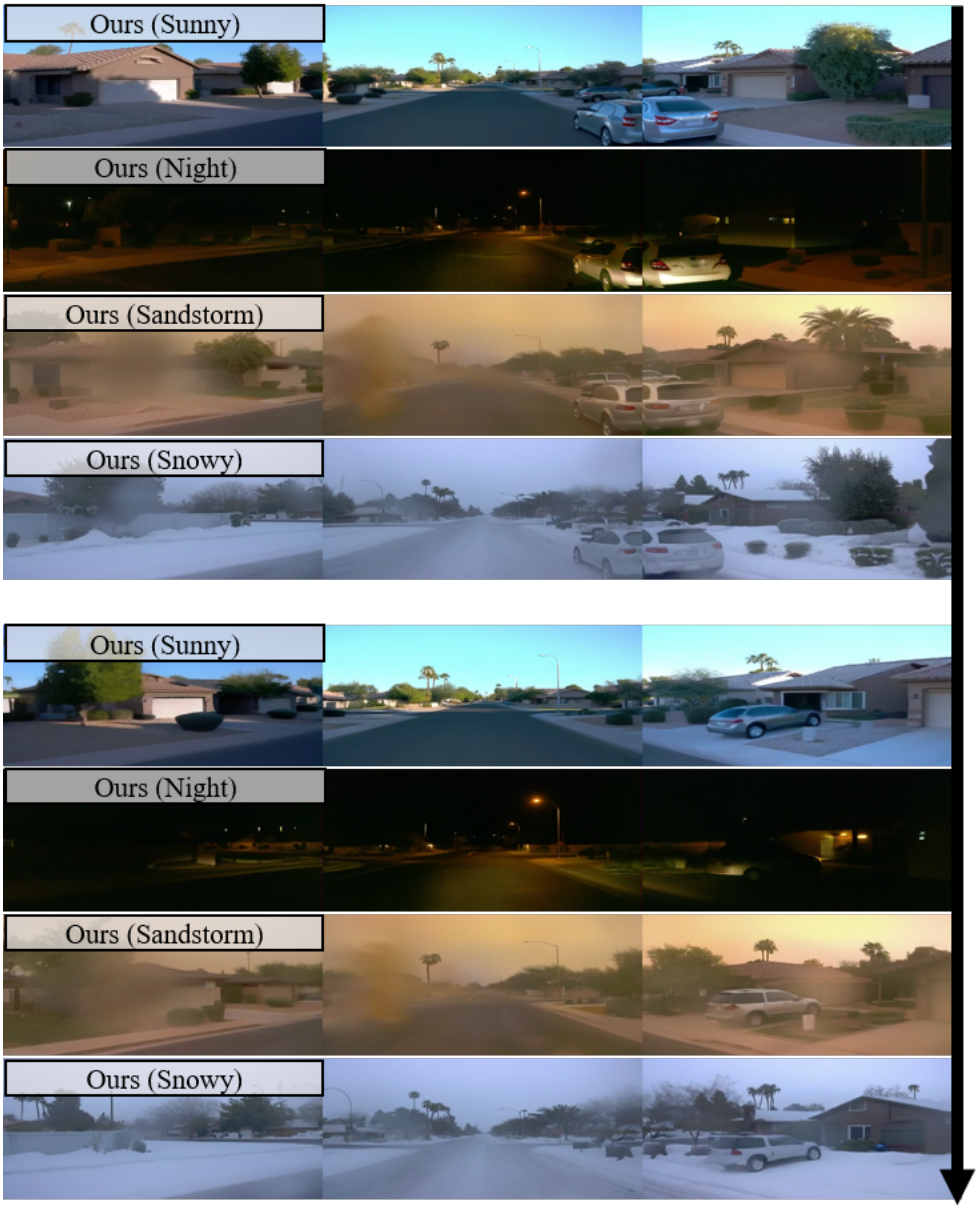

Figure 18. Additional visualizations of experimental results on Waymo.

Figure 19. Additional visualizations of our multi-view multi-modal generation results on Waymo (arbitrarily selected frames covering different time).