Appendix

A Qualitative Results

In this section, qualitative results are presented to demonstrate the performance of the proposed method.

A.1 Tokenizer Reconstructions

The visualization of tokenizer reconstructions are shown in Figure 7 and Figure 8. The proposed tokenizer can recover the image and Lidar with the unified BEV features.

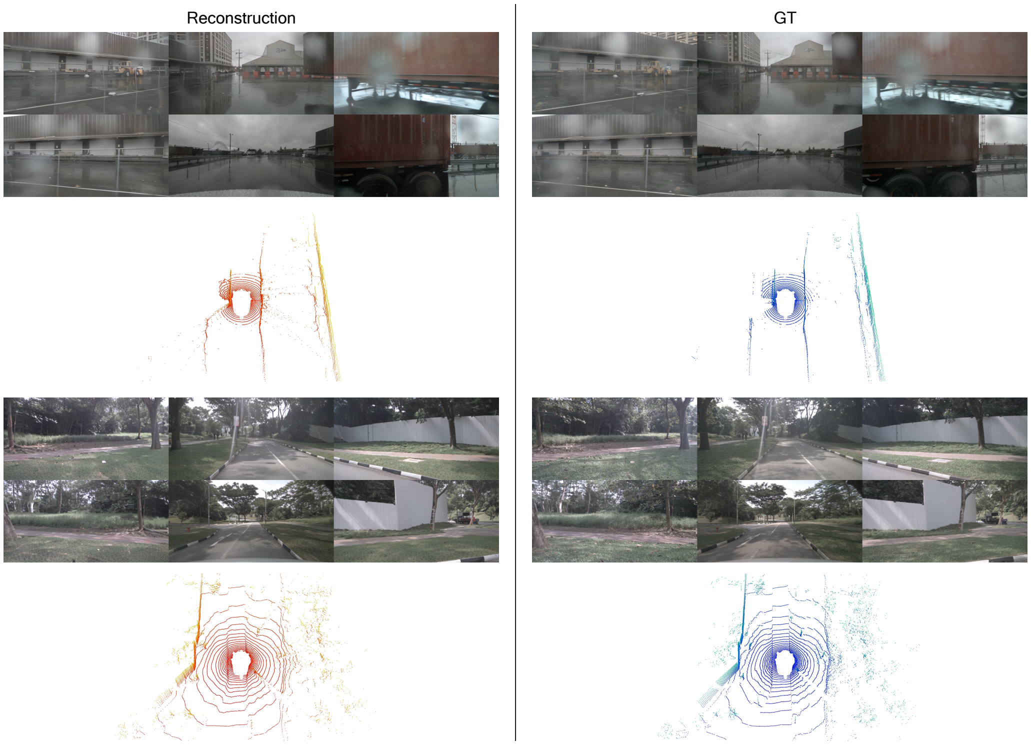

Figure 7: The visualization of LiDAR and video reconstructions on nuScenes dataset.

A.2 Multi-modal Future Predictions

Diverse generation. The proposed diffusion-based world model can produce high-quality future predictions with different driving conditions, and both the dynamic and static objects can be generated properly. The qualitative results are illustrated in Figure 9 and Figure 10.

Controllability. We present more visual results of controllability in Figure 11. The generated images and Lidar exhibit a high degree of consistency with action, which demonstrates that our world model has the potential of being a simulator.

PSNR metric. PSNR metric has the problem of being unable to differentiate between blurring and sharpening. As shown in Figure 12, the image quality of L & C is better the that of C, while the psnr metric of L & C is worse than that of C.

B Implementation Details

Training details of tokenizer. We trained our model using 32 GPUs, with a batch size of 1 per card. We used the AdamW optimizer with a learning rate of 5e-4, beta1=0.5, and beta2=0.9, following a

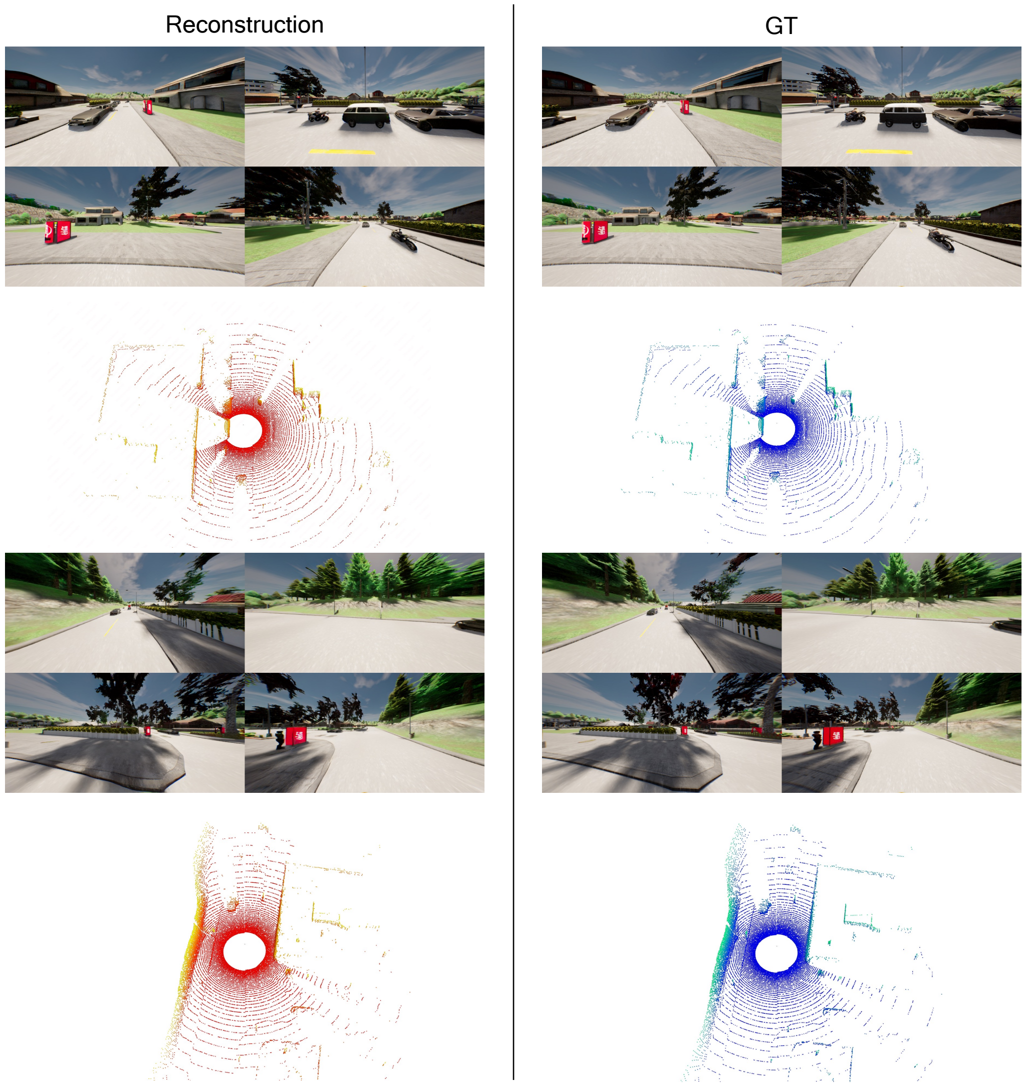

Figure 8: The visualization of LiDAR and video reconstructions on Carla dataset.

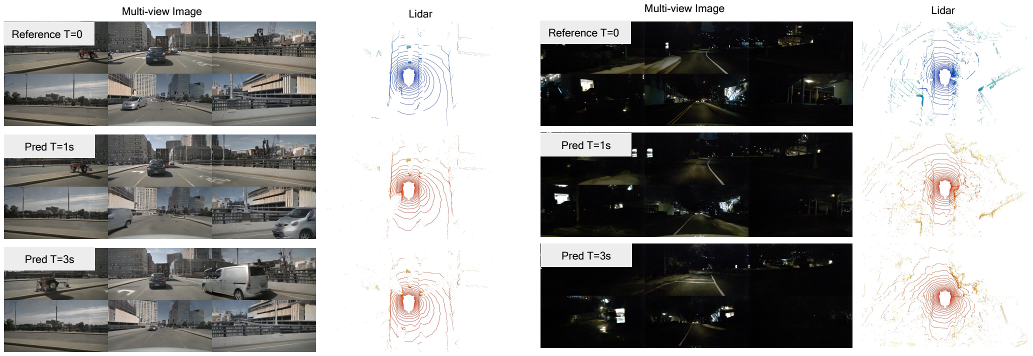

Figure 9: The visualization of LiDAR and future predictions on nuScenes dataset.

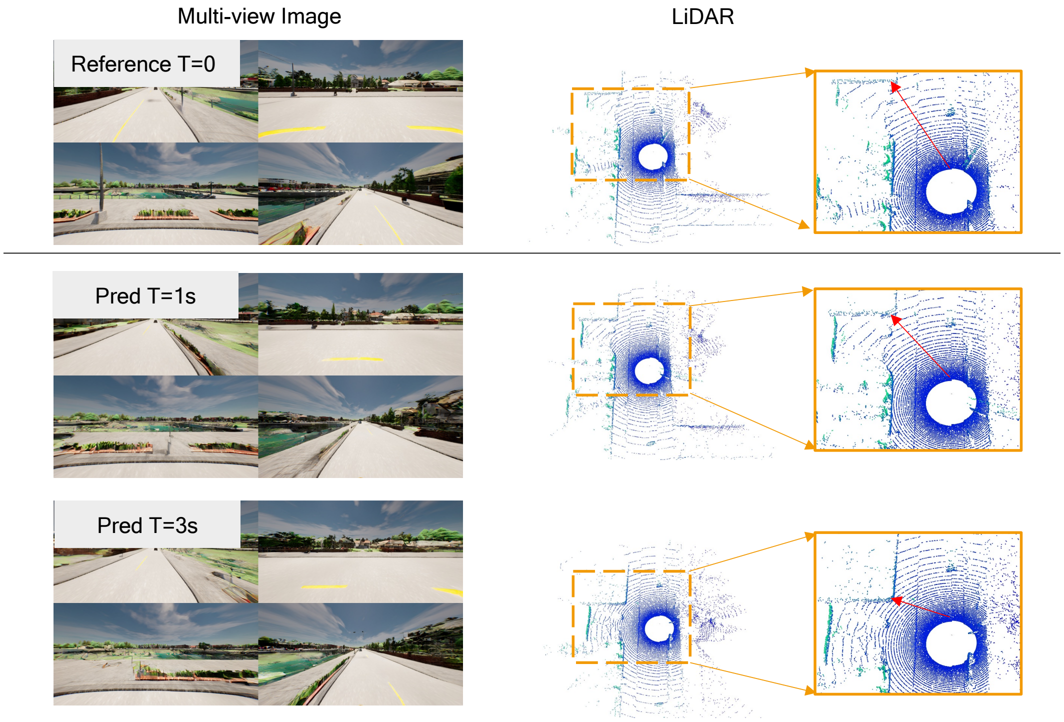

Figure 10: The visualization of LiDAR and future predictions on Carla dataset.

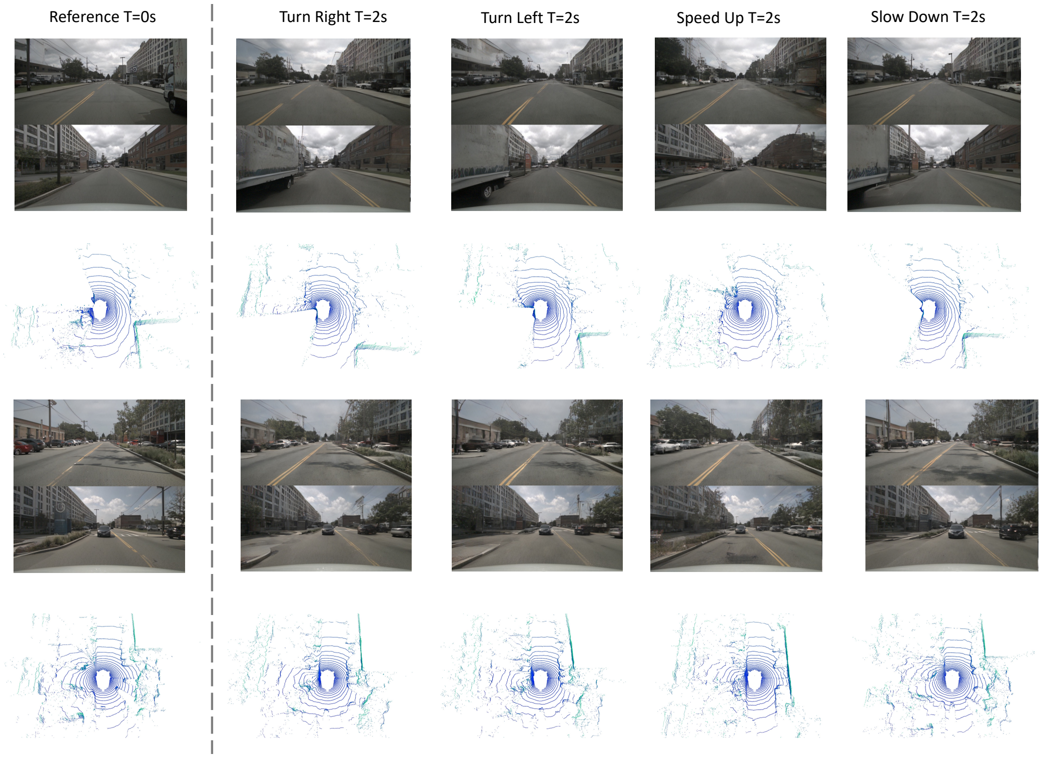

Figure 11: More visual results of controllability.

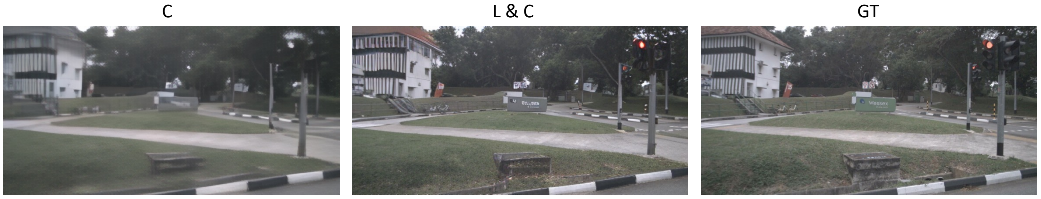

Figure 12: The visualization of C and L & C.

cosine learning rate decay strategy. The multi-task loss function includes a perceptual loss weight of 0.1, a lidar loss weight of 1.0, and an RGB L1 reconstruction loss weight of 1.0. For the GAN training, we employed a warm-up strategy, introducing the GAN loss after 30,000 iterations. The discriminator loss weight was set to 1.0, and the generator loss weight was set to 0.1.

Details on Upsampling from 2D BEV to 3D Voxel Features. The dimensional transformation proceeds as follows: (4, 96, 96) -> (256, 96, 96) via a linear layer -> (128, 192, 192) through Swin Blocks and upsampling (Patch Expanding in ViT-based methods) -> (128, 192, 192) again through additional Swin Blocks -> (4096, 192, 192) via a linear layer -> (16, 64, 384, 384) by reshaping, which represents the 3D voxel features. For the upsampling in Step 2, we adopt Patch Expanding, which is commonly used in ViT-based approaches and can be seen as the reverse operation of Patch Merging. The linear layer in Step 4 predicts a local region of shape (16, 64, r_y, r_x), where spatial sizes are adjusted (e.g., r_y=2, r_x=2), followed by reshaping in Step 5 to the final 3D feature shape.

Composition of 3D Voxel Features. Along each ray, we perform uniform sampling, and the depth t of the sampled points is a predefined value, not predicted by the model. The feature Vi at these sampled points is obtained through linear interpolation, while the blending weight w is predicted from the sampled features Vi (as described in Equation 1). This is a standard differentiable rendering process.

C Broader Impacts

The concept of a world model holds significant relevance and diverse applications within the realm of autonomous driving. It serves as a versatile tool, functioning as a simulator, a generator of long-tail data, and a pre-trained model for subsequent tasks. Our proposed method introduces a multi-modal BEV world model framework, designed to align seamlessly with the multi-sensor configurations inherent in existing autonomous driving models. Consequently, integrating our approach into current autonomous driving methodologies stands to yield substantial benefits.

D Limitations

It is widely acknowledged that inferring diffusion models typically demands around 50 steps to attain denoising results, a process characterized by its sluggishness and computational expense. Regrettably, we encounter similar challenges. As pioneers in the exploration of constructing a multi-modal world model, our primary emphasis lies on the generation quality within driving scenes, prioritizing it over computational overhead. Recognizing the significance of efficiency, we identify the adoption of one-step diffusion as a crucial direction for future improvement in the proposed method. Regarding the quality of the generated imagery, we have noticed that dynamic objects within the images sometimes suffer from blurriness. To address this and further improve their clarity and consistency, a dedicated module specifically tailored for dynamic objects may be necessary in the future.