DrivingGPT: Unifying Driving World Modeling and Planning with Multi-modal Autoregressive Transformers

Supplementary Material

A. Refining Video Generation with SVD Decoder.

Since independently decoding each frame-token to pixel space leads to temporally inconsistent video outputs, we further employ a video diffusion decoder [1] conditioned on frame-tokens to enable high-resolution and temporally consistent generation. As shown in Figure 6, we present four 32-second, 1024×576 examples from the NuPlan dataset.

During training, our video diffusion model is conditioned on image tokens generated by discretizing the input images. During inference, the model is conditioned on the predicted image tokens from the DrivingGPT.

During the inference, we first leverage the autoregressive nature of DrivingGPT for generating tokens for long video sequences beyond the training context length. We generate 64 frames in total for each video clip which spans 32 seconds. We predict tokens of the next 8 frames by conditioning on tokens of the last 8 frames of generated video and repeat the procedure until all 64 frames are generated. After obtaining tokens for the generated video clips, we use the VQ-VAE decoder [50] for converting the discrete tokens into continuous convolutional features. Since our autoregressive transformer models the driving video at a resolution of 288×512 while the SVD decoder models the driving video at a resolution of 576×1024, we upsample the convolutional features from the VQ-VAE decoder to align with the resolution of SVD. The convolutional features are dimensionally reduced to 4 dimensions using convolutional layers, which is then concatenated with the Gaussian noise as the conditional input to the denoising UNet. During the refinement stage, we decode 16 consecutive frames with the fine-tuned SVD decoder once, and then repeat this decoding process four times consecutively, ultimately generating a total video of 64 frames.

B. The Effect of Sampling Parameters on Video Generation.

As shown in Figure 7, we compare the impact of different sampling parameters, *top-k*, on generation. We find that the smaller the value of k, the smoother the image becomes; for example, the road appears flatter without cracks, and the shape of the car is smoother. Conversely, when k is larger, the image contains more detailed information; the cracks in the road are more pronounced, and the shape of the car is somewhat distorted.

C. Video Generation Samples under Different Driving Scenes.

We provide extra example video clips for demonstrating the generation diversity of our DrivingGPT on the project site `https://rogerchern.github.io/DrivingGPT`.

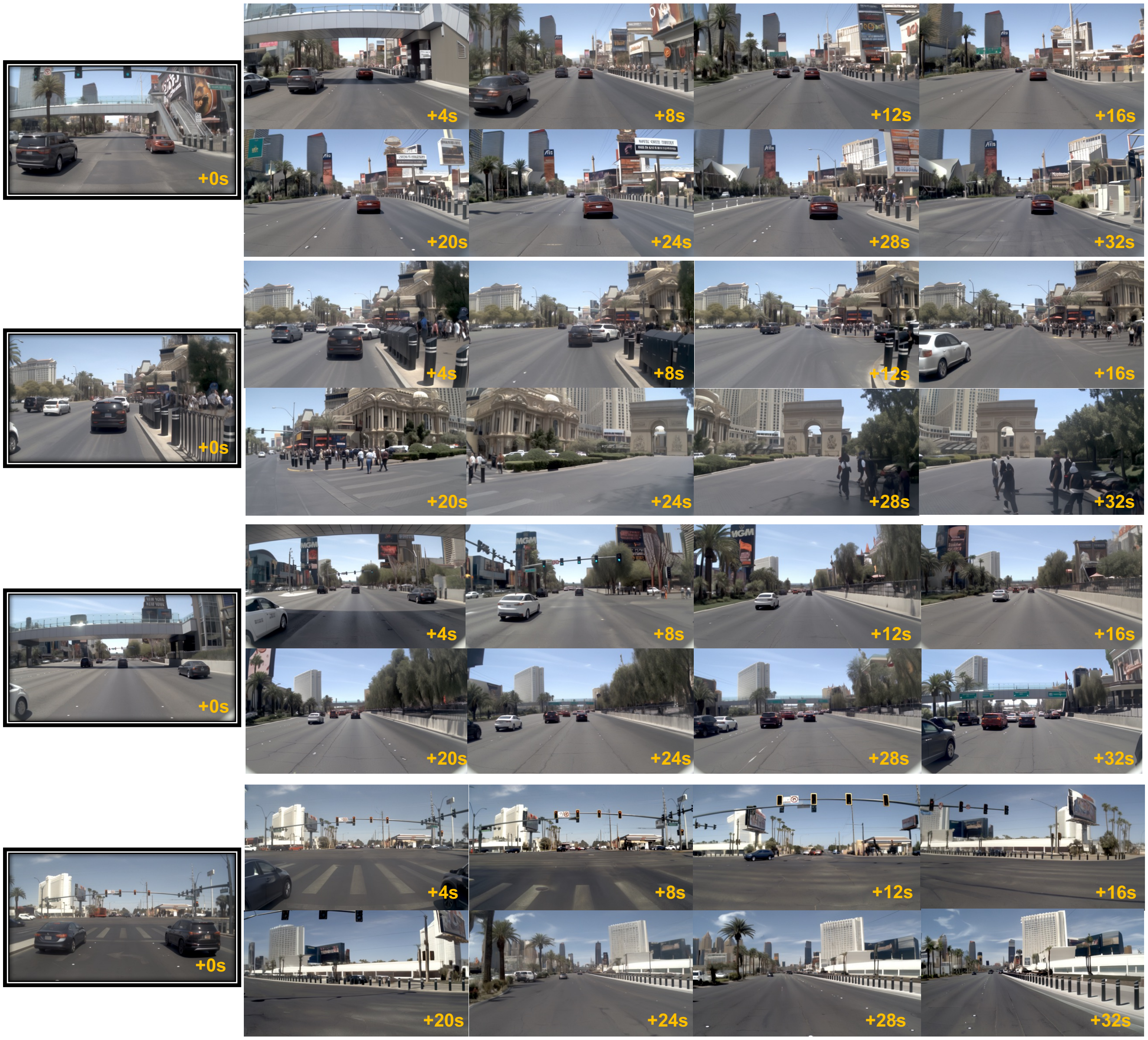

Figure 6. **Visualization of long video generation refined by SVD decoder**. We present four 32-second, 1024×576 examples from the NuPlan dataset. The predicted discrete tokens are of lower resolution, but by leveraging the SVD decoder, we can decode them to generate high-resolution outputs.

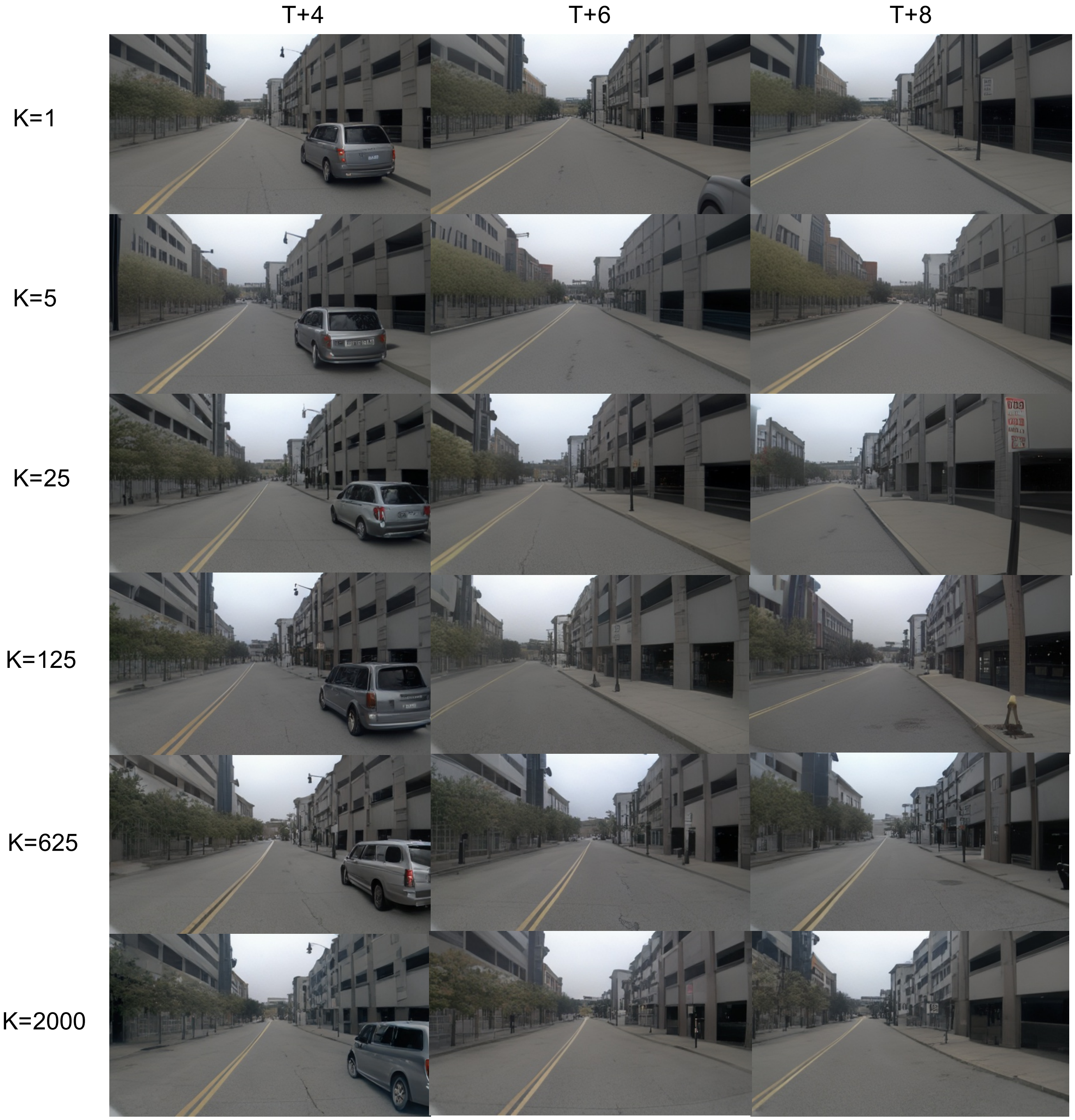

Figure 7. **Visualization of video generation under different sampling parameters**. we compare the impact of different sampling parameters, such as top-k, on generation. We find that the smaller the value of k, the smoother the image becomes; for example, the road appears flatter without cracks, and the shape of the car is smoother. Conversely, when k is larger, the image contains more detailed information; the cracks in the road are more pronounced, and the shape of the car is somewhat distorted.