Epona: Autoregressive Diffusion World Model for Autonomous Driving

Supplementary Material

A. Detailed Architecture of Dual-Single-Stream DiT

We are inspired by recent state-of-the-art image and video generation architectures [27, 33] and integrate dual-stream DiT blocks and single-stream DiT blocks to construct our TrajDiT and VisDiT. In the dual-stream DiT, condition information and noise are processed separately and interact only within the attention mechanism. In contrast, the single-stream DiT concatenates condition information and noise from the beginning for unified processing. Additionally, action control is mapped as an auxiliary control to obtain scale and shift parameters for adaptive modulation. The detailed architecture is illustrated in Fig. 9.

B. More Ablation Study

Effect of Temporal-aware DCAE Decoder. Considering that the original DCAE is an image-based autoencoder without temporal modeling capability, we incorporate a temporal interaction module before the DCAE decoder. As shown in Table 6, our world model achieves improved performance with the temporal module, effectively reducing flickering and enhancing the smoothness of the generated videos.

Table 6. **Comparison of the Generated Videos *w/* and *w/o* Temporal-aware DCAE Decoder Module on NuPlan [7] test set.** Temporal-aware DCAE Decoder can mitigate flickering artifacts and improve smoothness in generated videos.

Methods	$FVD_{10}\downarrow$	$FVD_{25}\downarrow$	$FVD_{40}\downarrow$
w/o Temporal Module	52.95	76.46	100.11
Ours	**50.77**	**61.46**	**74.88**

Effect of Different Context Length. We gradually increase the length of conditioned frames to investigate its impact on model performance. As shown in Table 7, as the number of conditioned frames increases, our world model improves in FVD performance due to longer historical information. However, longer conditioned frames require handling extended sequences, which poses computational challenges. Given our model setting, 10 frames represent the upper limit for conditioning. Therefore, we ultimately select 10 frames as the conditioning length in our approach.

C. More Discussions with Related Works

Comparison with GAIA-1 [24], DrivingGPT [11], and ADriver-I [28]. Compared to these multi-modal driving world models, our method adopts a fundamentally different architecture by directly integrating trajectory prediction into the video generation process via diffusion models. To the best of our knowledge, we are the *first* driving world model to use diffusion models for generating continuous, multi-step action trajectories, which brings two key advantages:

1. *Multi-step vs. single-step prediction.* Unlike prior approaches that interleave single-step image and action generation using transformers, our model predicts an entire N-step future trajectory in one shot. This is particularly beneficial for real-time motion planning in autonomous driving.
2. *Continuous vs. discrete action representation.* While existing methods discretize continuous action spaces into tokens, our diffusion model generates high-resolution continuous trajectories directly, enabling more precise planning and control.

Table 7. **Comparison of different condition frames on NuPlan [7] test set.** *Epona* generates better videos when conditioning more frames.

Frame number	$FVD_{10}\downarrow$	$FVD_{25}\downarrow$	$FVD_{40}\downarrow$
2	59.85	81.58	103.70
5	55.46	71.28	86.76
10	**50.77**	**61.46**	**74.88**

Among these methods, only DrivingGPT reports NAVISIM planning metrics. As shown in Tab. 8, our approach achieves significantly stronger results on this benchmark. Due to the absence of released code or full evaluation protocols for GAIA-1 and ADriver-I, we additionally compare against state-of-the-art end-to-end motion planners, where our model demonstrates competitive or superior performance.

Table 8. Comparison of planning results with DrivingGPT on the NAVSIM test set.

Method	**NC** ↑	**DAC** ↑	**TTC** ↑	**Comf.** ↑	**EP** ↑	**PDMS** ↑
DrivingGPT	**98.9**	90.7	**94.9**	95.6	79.7	82.4
Ours	97.9	**95.1**	93.8	**99.9**	**80.4**	**86.2**

Comparison with MagicDriveDiT [16] and InfinityDrive [20]. MagicDriveDiT and InfinityDrive are concurrent works focusing on video generation for autonomous driving. Based on their reported FVD scores on nuScenes (MagicDriveDiT: 94.84, InfinityDrive: 70.06), our method (82.83) exhibits competitive visual generation performance.

We acknowledge that MagicDriveDiT achieves slightly better visual quality, which we attribute to differences in

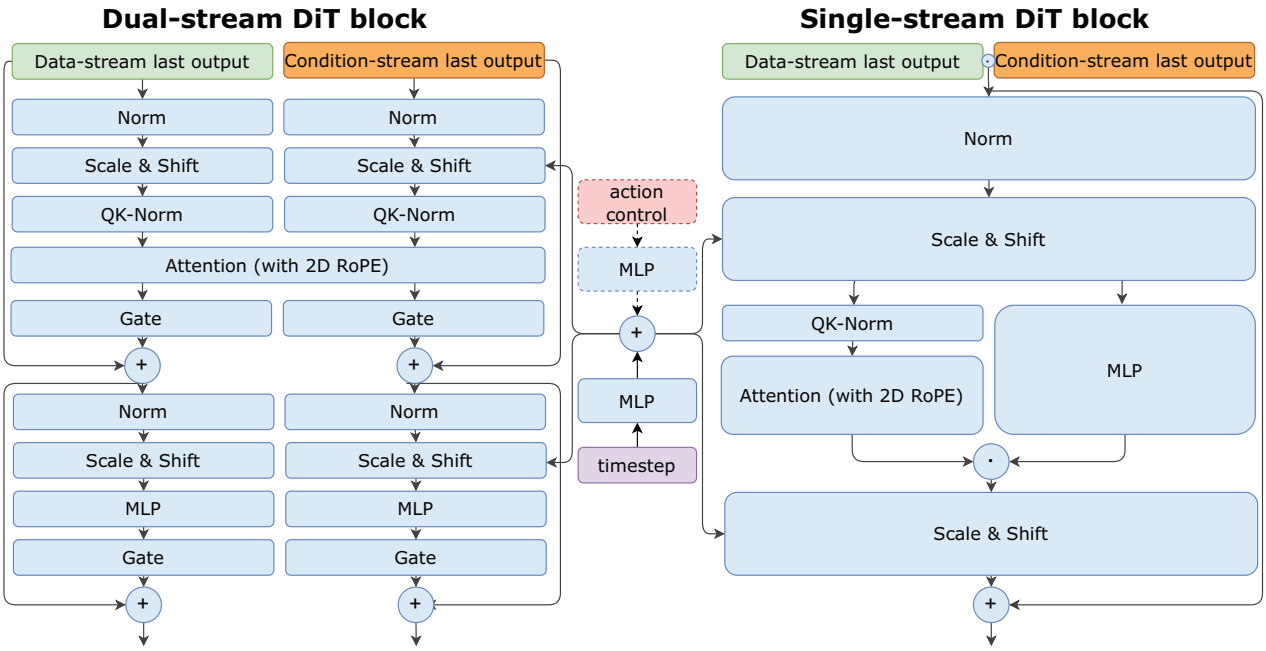

Figure 9. **Detailed architecture of dual-stram DiT and single stream DiT blocks.** We use nearly identical architecures for both TrajDiT and VisDiT, modified from text-image and video DiT architecture from [27, 33]. Action control is only for VisDiT.

video encoders: they utilize a specialized 3D-VAE, while we adopt a deep-compression autoencoder for better compression and training efficiency. This trade-off may introduce additional visual artifacts, and we plan to improve the DCAE component in future work.

More importantly, as elaborated in Sec. 3.2, these video diffusion-based methods are designed for scene synthesis without modeling causal dynamics or agent interactions. As a result, they lack support for flexible-length sequence generation and real-time planning, which are crucial in world model settings for decision making and policy learning.

Comparison with Transfusion [71] and JanusFlow [40]. While these multimodal generative models also combine diffusion and autoregression, their design principles differ significantly from ours. Transfusion and JanusFlow combine *token-wise text autoregression* and diffusion for image understanding and generation. In contrast, our model combine *frame-wise latent autoregression* and diffusion with novel decoupled architecture design to tackle the unique problem of *temporal dynamics and coherence* with video inputs and outputs, which is more challenging.

D. More Long-term Video Generation Results

As shown in Fig. 10, we present the generation of minute-long ultra-long videos while maintaining high-quality visuals and preserving the integrity and details of surrounding buildings and vehicles. Additionally, our world model continuously generates the next frames with new contents without experiencing context drift.

E. Limitation and Future Work

Despite Epona's unified framework for long-horizon generation, controllable simulation, and real-time planning, the model still falls short in accurately capturing physical dynamics and achieving high-fidelity visual synthesis. Future work will explore integrating stronger inductive biases—such as 3D structure, multi-view geometry, and multimodal information—to improve physical modeling and interaction understanding. We also aim to push the boundaries of visual quality and diversity by scaling training to larger datasets, moving toward more generalizable and physically grounded world models.

Figure 10. Visualization of Longer Videos. Our world model is capable of generating extended videos (140 seconds) while maintaining high visual quality and detailed vehicles and buildings.