See Tomorrow, Act Today: Foresight-Driven Autonomous Driving Supplementary Material

Bozhou Zhang[1,2*], Nan Song[1,2*], Yuang Wang[1], Jiankang Deng[3], Xiatian Zhu[4], Li Zhang[1,2‡]

[1]School of Data Science, Fudan University [2]Shanghai Innovation Institute

[3]Imperial College London [4]University of Surrey

1. Discussions

Discussion 1: More discussions with related work.

World model:

Besides the comparisons presented in the main paper, we further discuss how ForeSight differs from methods such as Epona [24] and DrivingGPT [4], highlighting the unique contributions of our approach.

Although Epona, DrivingGPT, and our model all generate future frames and plan future trajectories, their focuses are fundamentally different. Epona and DrivingGPT are primarily world-modeling methods—their core contribution lies in training generative models for future-scene synthesis, whereas trajectory planning is treated as an auxiliary output. DrivingGPT uses discrete tokens to generate future trajectories, while Epona employs a diffusion transformer for trajectory generation. Both methods jointly generate future frames and trajectories within a unified generative framework.

In contrast, our approach follows a world–action design tailored specifically for end-to-end autonomous driving. ForeSight leverages a foundation world model as a unified module for perception, comprehension, and future-scene imagination, while a specially designed action module is optimized to produce high-quality trajectories. This separation allows the system to fully exploit future predictions while maintaining strong planning performance.

Planning model:

Recent planning models explore diverse directions to enhance planning performance, including diffusion-based decoders [9, 11, 12, 14, 15, 18, 25], reinforcement-learning–based approaches [3, 10, 13, 20, 22], test-time training [23], clustered anchored trajectory priors [16, 17], mixture-of-experts architectures [6], and Gaussian-feature–fusion strategies [19]. All of these methods demonstrate strong empirical performance.

These approaches are largely orthogonal to ours: ForeSight intentionally adopts simple action-decoding modules so that our core contribution—the integration of a foun-

*Equal contribution, ‡Corresponding author (lizhangfd@fudan.edu.cn).

Table 1. Generation performance comparison.

Method	FVD_{10}
Epona [24]	50.77
ForeSight (Ours)	54.63

dation world model to guide planning—is clearly isolated and easy to plug into existing systems. We believe that combining these advanced planning techniques with our world–action framework would likely yield even better performance.

Discussion 2: About the world model architecture.

Our ForeSight relies on the generated future-frame features for planning, which means it is not restricted to any specific world-model architecture, such as diffusion-based models [7, 24] or GPT-based models [4, 8]. In our main experiments, we adopt Epona [24] as the primary world model, considering both its capability and open-source availability. Nevertheless, we also provide results using alternative world-model architectures in Section 2.2, demonstrating that ForeSight is compatible with diverse backbone designs.

Discussion 3: About the current encoder.

In autonomous driving, accurate trajectory planning often requires a detailed understanding of the surroundings, typically obtained from multi-view cameras or LiDAR point clouds. To provide this spatial awareness, we incorporate a lightweight encoder based on TransFuser [21] to extract current-frame features as an additional supplement. However, this component is not strictly necessary. As shown in Section 2.2, we also report results without the current encoder. With future advances in world models—particularly in high-resolution and multi-view generation—we expect that the current encoder can eventually be removed and fully replaced by a unified world–action framework.

Discussion 4: Efficiency analysis.

For the parameter size, ForeSight mainly consists of three components: the foundation world model, the current encoder, and the action decoder. For the foundation world model, we adopt Epona [24], which contains 2.5 B parameters. The current encoder is largely inherited from TransFuser [21], comprising 52 M parameters. The action decoder contains an additional 21 M parameters.

For inference time, we evaluate our model on an NVIDIA H100 GPU. The average inference time of ForeSight is 900 ms, with the majority (approximately 870 ms) attributed to the world model [24]. As world-model architectures continue to advance, their inference efficiency is expected to improve substantially, making the overall system considerably more deployment-friendly.

Table 2. Performance comparison with and without the current encoder.

	DAC ↑	TTC ↑	EP ↑	PDMS ↑
w/o Current	96.3	95.4	81.7	88.2
ForeSight	97.2	94.8	83.5	89.3

2. Experiments

2.1. Implementation details

Besides the implementation details provided in the main paper, additional information is included here to ensure full reproducibility. For the foundation world model, Epona [24] is adopted. Its native generation frequency is 5 Hz, whereas the planning frequency in our system is 2 Hz. To match this temporal resolution, Epona is first finetuned on the nuPlan [2] dataset at 2 Hz and subsequently frozen when training the full pipeline on NAVSIM [5]. For the current encoder, the design largely follows TransFuser [21]. During training, the current encoder and the action decoder are first pretrained without the future feature cross-attention or the WM-QFormer modules. Afterwards, the full model is trained end-to-end with all components enabled, except that the world model remains frozen.

2.2. More experiment results

Generation performance of the world model. As shown in Table 1, we report the Fréchet Video Distance (FVD) of ForeSight and Epona [24] on the nuPlan [2] dataset. The results indicate that ForeSight retains nearly the same generation capability as Epona after finetuning.

Performance without the current encoder. As shown in Table 2, we report the performance of our model without the current encoder on the NAVSIM [5] dataset. The results show that the model still achieves strong performance even when this module is removed, demonstrating that it is not strictly necessary. Nevertheless, the current encoder is retained in the full system, as it further enhances robustness and overall capability.

Performance with an alternative world-model architecture. As shown in Table 3, we also evaluate our framework using Vista [7] as the foundation world model. Since Vista is trained on the nuScenes [1] dataset, the experiments are conducted on this dataset for the end-to-end planning task. The results indicate that ForeSight with Vista achieves strong performance as well, demonstrating that our framework is not restricted to a specific world-model architecture.

Table 3. Performance with different world-model architectures on the nuScenes dataset for the planning task.

Method	L2 (m) ↓				Col. Rate (%) ↓			
	$1s$	$2s$	$3s$	Avg.	$1s$	$2s$	$3s$	Avg.
ForeSight-Vista	0.42	0.63	0.88	0.64	0.08	0.22	0.51	0.27
ForeSight-Epona	0.36	0.55	0.93	0.62	0.04	0.12	0.37	0.18

3. Qualitative results

As shown in Figure 1, additional qualitative results of ForeSight are provided, including turning behaviors (parts (a) and (c)), a traffic-congestion scenario (part (b)), and a fast-driving behavior (part (d)). Across all cases, the model not only predicts future frames accurately but also produces precise future trajectories.

4. Failure cases

Although ForeSight is powerful, it still fails in certain scenarios. Representative failure cases are shown in Figure 2, which may provide insights for future research.

In part (a), the scenario involves a right-turning maneuver. The foundation world model accurately predicts both the turning motion and the post-turn scene. However, the action decoder generates an overly conservative and slow trajectory. This indicates that the world model and the action model should be more tightly coupled so that the planner can better leverage future predictions for trajectory generation.

In part (b), the scenario involves fast driving over a long distance within the planning horizon, and the road is highly winding. While the foundation world model produces accurate predictions initially, it fails in the later stages due to the increasing curvature of the road. This suggests that long-range prediction capability remains a challenge for the world model. Nevertheless, the action model still produces an accurate trajectory. This highlights the importance of using current-frame features as an additional supplement, which significantly enhances the overall robustness of the system.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2

[2] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint*, 2021. 2

[3] Keyu Chen, Wenchao Sun, Hao Cheng, and Sifa Zheng. Rift: Closed-loop rl fine-tuning for realistic and controllable traffic simulation. *arXiv preprint*, 2025. 1

[4] Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *ICCV*, 2025. 1, 2

[5] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *NeurIPS*, 2024. 2, 4, 5

[6] Renju Feng, Ning Xi, Duanfeng Chu, Rukang Wang, Zejian Deng, Anzheng Wang, Liping Lu, Jinxiang Wang, and Yanjun Huang. Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving. *arXiv preprint*, 2025. 1

[7] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *NeurIPS*, 2024. 2

[8] Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructing world model for autonomous driving via video gpt. *arXiv preprint*, 2024. 2

[9] Anqing Jiang, Yu Gao, Zhigang Sun, Yiru Wang, Jijun Wang, Jinghao Chai, Qian Cao, Yuweng Heng, Hao Jiang, Yunda Dong, et al. Diffvla: Vision-language guided diffusion planning for autonomous driving. *arXiv preprint*, 2025. 1

[10] Anqing Jiang, Yu Gao, Yiru Wang, Zhigang Sun, Shuo Wang, Yuwen Heng, Hao Sun, Shichen Tang, Lijuan Zhu, Jinhao Chai, et al. Irl-vla: Training an vision-language-action policy via reward world model. *arXiv preprint*, 2025. 1

[11] Hao Jiang, Zhipeng Zhang, Yu Gao, Zhigang Sun, Yiru Wang, Yuwen Heng, Shuo Wang, Jinhao Chai, Zhuo Chen, Hao Zhao, et al. Flowdrive: Energy flow field for end-to-end autonomous driving. *arXiv preprint*, 2025. 1

[12] Xuefeng Jiang, Yuan Ma, Pengxiang Li, Leimeng Xu, Xin Wen, Kun Zhan, Zhongpu Xia, Peng Jia, XianPeng Lang, and Sheng Sun. Transdiffuser: End-to-end trajectory generation with decorrelated multi-modal representation for autonomous driving. *arXiv preprint*, 2025. 1

[13] Siwen Jiao, Kangan Qian, Hao Ye, Yang Zhong, Ziang Luo, Sicong Jiang, Zilin Huang, Yangyi Fang, Jinyu Miao, Zheng Fu, et al. Evadrive: Evolutionary adversarial policy optimization for end-to-end autonomous driving. *arXiv preprint*, 2025. 1

[14] Pengxiang Li, Yinan Zheng, Yue Wang, Huimin Wang, Hang

Figure 1. Visualization of **ForeSight** on the NAVSIM [5] dataset. The left panel shows the planned trajectories in the BEV view, while the right panel presents the generated future video over the next 8 time steps. The ground-truth trajectory is depicted in green, and the final planned trajectory is highlighted in orange.

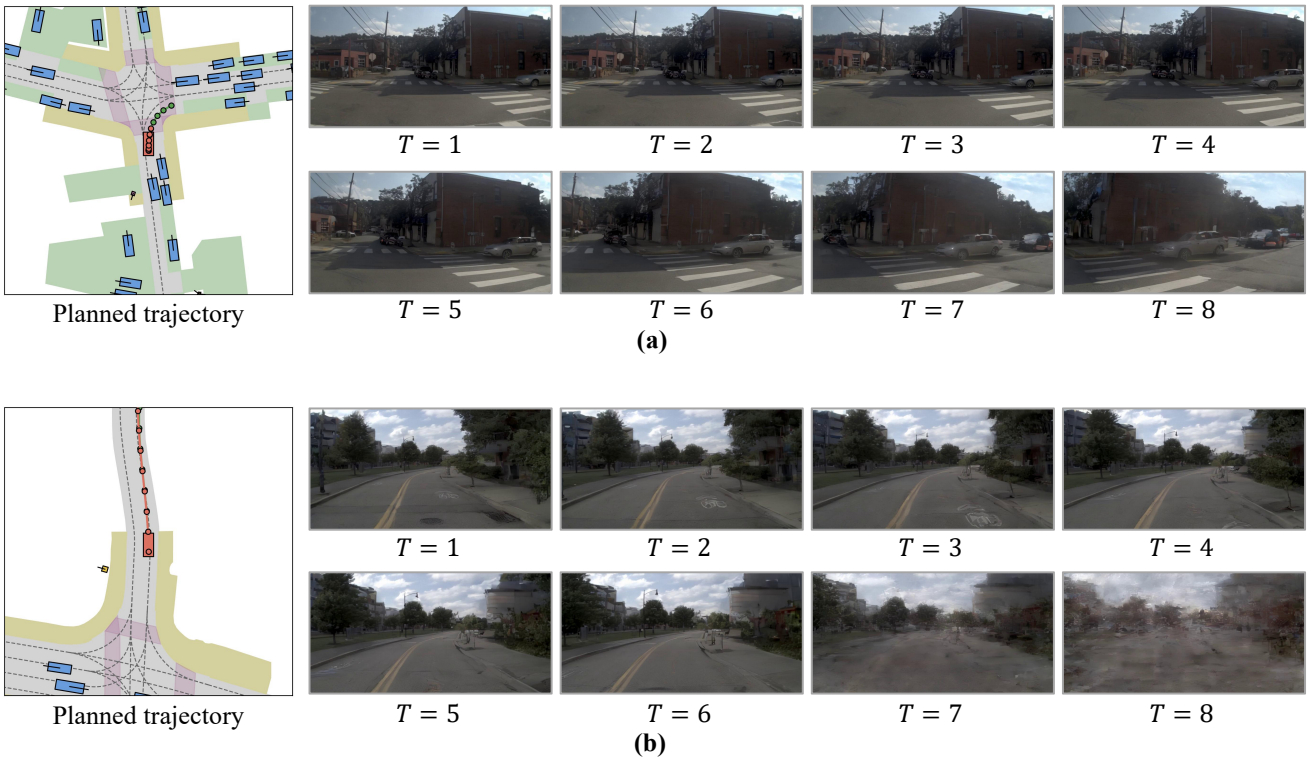

Figure 2. Visualization of *failure cases* for **ForeSight** on the NAVSIM [5] dataset. The left panel shows the planned trajectories in the BEV view, while the right panel presents the generated future video over the next 8 time steps. The ground-truth trajectory is depicted in green, and the final planned trajectory is highlighted in orange.

Zhao, Jingjing Liu, Xianyuan Zhan, Kun Zhan, and Xianpeng Lang. Discrete diffusion for reflective vision-language-action models in autonomous driving. *arXiv preprint*, 2025. 1

[15] Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint*, 2025. 1

[16] Zhenxin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Zuxuan Wu, and Jose M Alvarez. Hydra-next: Robust closed-loop driving with open-loop training. *arXiv preprint*, 2025. 1

[17] Zhenxin Li, Wenhao Yao, Zi Wang, Xinglong Sun, Joshua Chen, Nadine Chang, Maying Shen, Zuxuan Wu, Shiyi Lan, and Jose M Alvarez. Generalized trajectory scoring for end-to-end multimodal planning. *arXiv preprint*, 2025. 1

[18] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *CVPR*, 2025. 1

[19] Shuai Liu, Quanmin Liang, Zefeng Li, Boyang Li, and Kai Huang. Gaussianfusion: Gaussian-based multi-sensor fusion for end-to-end autonomous driving. *arXiv preprint*, 2025. 1

[20] Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, et al. Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving. *arXiv preprint*, 2025. 1

[21] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, 2021. 2

[22] Shuyao Shang, Yuntao Chen, Yuqi Wang, Yingyan Li, and Zhaoxiang Zhang. Drivedpo: Policy learning via safety dpo for end-to-end autonomous driving. In *NeurIPS*, 2025. 1

[23] Chonghao Sima, Kashyap Chitta, Zhiding Yu, Shiyi Lan, Ping Luo, Andreas Geiger, Hongyang Li, and Jose M Alvarez. Centaur: Robust end-to-end autonomous driving with test-time training. *arXiv preprint*, 2025. 1

[24] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *ICCV*, 2025. 1, 2

[25] Rui Zhao, Yuze Fan, Ziguo Chen, Fei Gao, and Zhenhai Gao. Diffe2e: Rethinking end-to-end driving with a hybrid action diffusion and supervised policy. *arXiv preprint*, 2025. 1