Other Vehicle Trajectories Are Also Needed: A Driving World Model Unifies Ego-Other Vehicle Trajectories in Video Latent Space

Jian Zhu*, Zhengyu Jia, Tian Gao, Jiaxin Deng, Shidi Li, Lang Zhang†, Fu Liu, Peng Jia, Xianpeng Lang

Li Auto Inc.

Abstract

Advanced end-to-end autonomous driving systems predict other vehicles' motions and plan ego vehicle's trajectory. The world model that can foresee the outcome of the trajectory has been used to evaluate the autonomous driving system. However, existing world models predominantly emphasize the trajectory of the ego vehicle and leave other vehicles uncontrollable. This limitation hinders their ability to realistically simulate the interaction between the ego vehicle and the driving scenario. In this paper, we propose a driving **W**orld **M**odel named EOT-WM, unifying **E**go-**O**ther vehicle **T**rajectories in videos for driving simulation. Specifically, it remains a challenge to match multiple trajectories in the BEV space with each vehicle in the video to control the video generation. We first project ego-other vehicle trajectories in the BEV space into the image coordinate for vehicle-trajectory match via pixel positions. Then, trajectory videos are encoded by the Spatial-Temporal Variational Auto Encoder to align with driving video latents spatially and temporally in the unified visual space. A trajectory-injected diffusion Transformer is further designed to denoise the noisy video latents for video generation with the guidance of ego-other vehicle trajectories. In addition, we propose a metric based on control latent similarity to evaluate the controllability of trajectories. Extensive experiments are conducted on the nuScenes dataset, and the proposed model outperforms the state-of-the-art method by 30% in FID and 55% in FVD. The model can also predict unseen driving scenes with self-produced trajectories.

Introduction

End-to-end autonomous driving (Hu et al. 2023; Tian et al. 2024; Li et al. 2025) has gained increasing attention recently, since the approach integrates all modules (e.g. perception, decision, and planning modules) into a model optimized jointly to directly output planning results based on input multi-sensor data. Despite the promising performance of the end-to-end autonomous driving model, effectively handling out-of-distribution scenarios continues to be a significant challenge, particularly since such situations are often hazardous and costly to simulate. The world model (Wang et al. 2023b, 2024b; Gao et al. 2024; Hassan et al. 2025) can predict future driving scenes based on historical observations and future driving actions, which is a potential solution to evaluate the autonomous driving model and avoid catastrophic errors.

Some world models (Gao et al. 2024; Yang et al. 2024a) explore predicting future driving scenes with the trajectory, since it can reflect the driving action of the vehicle in the scene more precisely. Despite advanced end-to-end autonomous driving systems (e.g. VAD (Jiang et al. 2023)) can predict other vehicles' motions and plan ego vehicle's trajectory, these world models primarily focus on the ego vehicle trajectory. They view the trajectory as a series of points in the bird's eye view (BEV) space (Hu et al. 2021; Li et al. 2024), which are directly encoded as the condition for video generation. However, there are three main drawbacks in the approaches mentioned above. Firstly, they just consider the ego vehicle trajectory and leave other vehicles uncontrollable in the generated video. As a result, the model cannot realistically simulate the interaction between the ego vehicle and the driving scenario, and generate diverse novel scenes by changing other vehicle trajectories as well. Secondly, the distribution of the encoding based on the trajectory points in the BEV space is quite different from that of the video latents in visual modality without aligning their feature space. In addition, it is impractical to correspond multiple trajectories in the BEV space to numerous vehicles in visual modality, since the BEV space is mismatched with the video space.

To tackle above issue, a driving world model unifying ego-other vehicle trajectories in videos named EOT-WM is proposed in this paper. As is shown in Fig. 1, Vista (Gao et al. 2024) only uses the ego vehicle trajectory, and the stationary other vehicle in the groundtruth video also moves forward in the generated video. The proposed EOT-WM can generate more realistic videos with controllable ego and other vehicle actions. In addition, the model can also generate video matched based on novel self-produced trajectories. Specifically, instead of representing the trajectory via points in the BEV space, we project these points into the image coordinate and plot the ego vehicle trajectory and the other vehicle trajectories in separate blank videos to generate trajectory videos for learning in unified visual modality. Then, we adopt Spatial-Temporal Variational Auto Encoder (STVAE) to encode the scene video and trajectory video to achieve scene video latents and trajectory latents

*Corresponding Author (jianzhu823@gmail.com)

†Project Leader

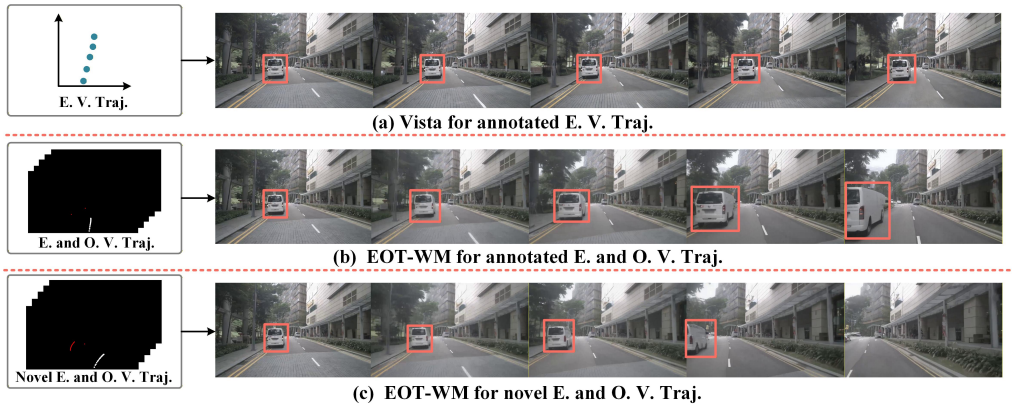

Figure 1: The proposed EOT-WM is capable of generating more realistic videos with controllable ego and other vehicle trajectories. These trajectories are represented in video space for EOT-WM instead of BEV space for previous works such as Vista. E. V. Traj. and O. V. Traj. denote ego and other vehicle trajectories, respectively. Novel trajectory means self-produced trajectory not included in the dataset.

with sharing feature space. Moreover, the scene and trajectory latents achieved in this manner are aligned temporally and spatially to realize effective control. Finally, we design Trajectory-injected Diffusion Transformer (TiDiT) to integrate motion guidances provided by trajectory latents into video latents for denoising the noisy video latents more precisely. As a result, the entire model can predict future frames based on the given initial frames with the control of text and trajectory. To evaluate the controllability of trajectories, we propose a metric based on control latent similarity, which compares the predicted trajectory latents with the groundtruth trajectory latents.

The main contributions of the paper can be summarized as follows.

- We firstly propose a driving world model with ego and other vehicle trajectories, more realistically simulating the interaction between the ego vehicle and the driving scenario and able to generate diverse novel scenes with alterable ego-other trajectories.
- We propose to represent the trajectory as the video and encode the trajectory video via the driving video encoder to make each trajectory aligned with each vehicle in the unified visual space.
- A trajectory-injected diffusion Transformer is designed to denoise the noisy video latents more precisely via the ego-other vehicle trajectories. A metric based on control latent similarity is further proposed to evaluate the controllability of trajectories.

Related Work

Video Generation

Video generation is capable of understanding the world and generating realistic video samples. Various kinds of models have been studied in the past, including VAE-based (Villegas et al. 2019; Franceschi et al. 2020), flow-based (Dorkenwald et al. 2021; Kumar et al. 2020), GAN-based (Brooks et al. 2022; Yu et al. 2022) and auto-regressive models (Ge et al. 2022; Weissenborn, Täckström, and Uszkoreit 2020). Recently, diffusion models have achieved breakthroughs in image generation (Nichol et al. 2022; Rombach et al. 2022), and diffusion models are also introduced into video generation (Blattmann et al. 2023b; Guo et al. 2024; Ho et al. 2022; Yang et al. 2024b). Most of these models (Blattmann et al. 2023b; Guo et al. 2024; Ho et al. 2022; Yang et al. 2024b) use the text as the condition to control video generation. To explore more condition forms for controlling video generation, image-to-video generative models (Blattmann et al. 2023a; Chen et al. 2023) are studied since the image can provide more specific priors. However, above models (Blattmann et al. 2023b; Guo et al. 2024; Ho et al. 2022; Yang et al. 2024b; Blattmann et al. 2023a; Chen et al. 2023) are incapable of predicting the future state, which are crucial in autonomous driving. In addition, camera motion (Wang et al. 2024c) and object motion (Wang et al. 2024c; Zhang et al. 2024) are employed to realize more flexible video generation. Unfortunately, these models are not specifically designed to tackle complicated autonomous driving videos and cannot model vehicle actions well. In this paper, we propose EOT-WM that jointly uses texts and trajectories to predict future video frames based on given initial states.

World Model

The world model can predict the future states based on historical observations and future actions, which has been extensively applied in simulated games (Ha and Schmidhuber 2018; Hafner et al. 2020, 2021) and indoor embodiment (Koh et al. 2021; Wang et al. 2023a; Mendonca, Bahl, and Pathak 2023). Recently, the world model has attracted great attention in autonomous driving (Wang et al. 2023b; Lu et al. 2024; Wang et al. 2024b; Yang et al. 2024a; Gao et al. 2024; Hassan et al. 2025; Agarwal et al. 2025). WoVoGen (Lu et al. 2024) and Drive-WM (Wang et al. 2024b) are devoted to generating 6-view driving videos based on the given conditions. Drivedreamer (Wang et al. 2023b), GenAD (Yang et al. 2024a), Vista (Gao et al. 2024), and

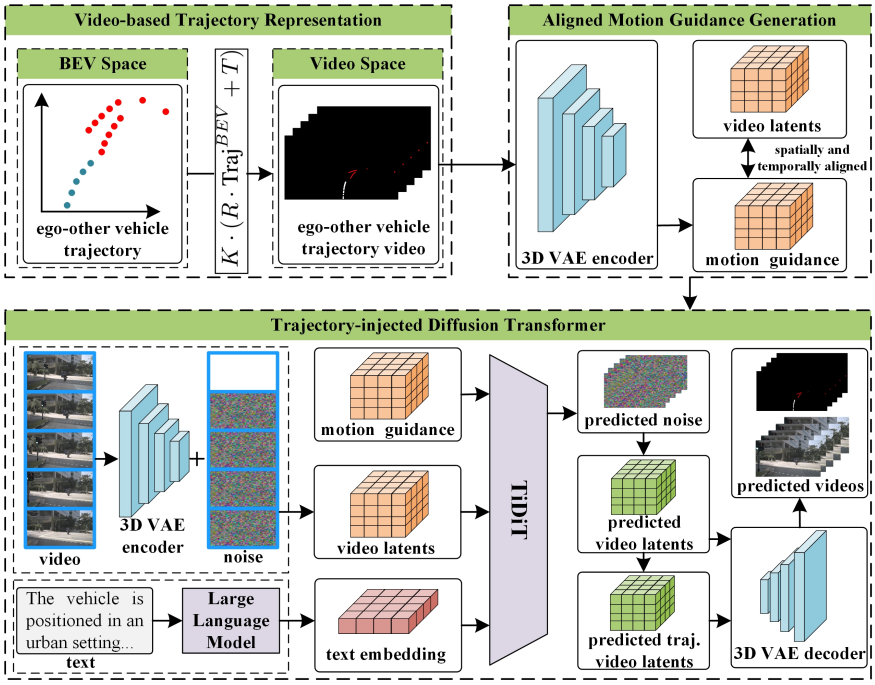

Figure 2: Illustration of the proposed EOT-WM.

GEM (Hassan et al. 2025) develop world models to predict future driving scenes via the front-view video. However, most world models mentioned above concentrate on the actions of the ego vehicle and ignore those of the other vehicles, making the interaction with the environment insufficient in these world models. Although GEM (Hassan et al. 2025) uses future object features and human poses to generate future videos, it is more impractical for autonomous driving systems to obtain them compared with future trajectories. Particularly, Cosmos (Agarwal et al. 2025) is a commercial model comprising over 7 billion parameters, which imposes substantial computational overhead during both training and inference. In this paper, the proposed EOT-WM uses the ego-other vehicle trajectories to represent the motions of vehicles in the scene, and generate more realistic videos with all vehicles controllable.

EOT-WM Framework

The overall architecture of the proposed EOT-WM framework is illustrated in Fig. 2. We build EOT-WM based on CogvideoX (Yang et al. 2024b), and modify it as the world model architecture with injected trajectory condition. Specifically, the proposed EOT-WM framework consists of Video-based Trajectory Representation (VTR), Aligned Motion Guidance Generation (AMGG), and Trajectory-injected Diffusion Transformer (TiDiT), which are described in detail in the following sections.

Video-based Trajectory Representation

The trajectory in the end-to-end autonomous driving system is usually represented as a series of points in the BEV space for planning, which can be formulated as $\text{Traj}^{BEV} = \{(x_1, y_1), (x_2, y_2), ..., (x_T, y_T)\}$ and T is the number of points. However, the world model in this paper aims to generate a video that simulates the driving scene based on vehicle trajectories. All elements in the generated video such as vehicles and roads are represented by pixels, leading to the natural mismatch between the points in the BEV space and the pixels in the video. In addition, it is impractical to correspond multiple trajectories in the BEV space to numerous vehicles in the video. Therefore, we propose to represent the trajectory in the form of video to act as the condition.

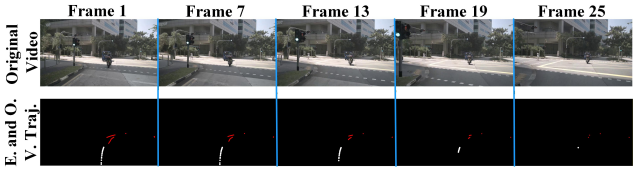

Figure 3: Illustration of the original video, other vehicle trajectory (O. V. Traj.) and ego vehicle trajectory (E. V. Traj.) used for the proposed EOT-WM. To be brief, we only visualize the 1st, 7th, 13th, 19th, 25th frames.

Given T trajectory points in the BEV space and T video frames, each trajectory point is projected into the image coordinate of the first frame based on the camera extrinsic and intrinsic parameters. The obtained points are denoted as $\text{Traj}^I = \{(x_1^I, y_1^I), (x_2^I, y_2^I), ..., (x_T^I, y_T^I)\}$. The projection can be calculated as

$$\text{Traj}^I = K \cdot (R \cdot \text{Traj}^{BEV} + B), \quad (1)$$

where R and B are rotation matrix and translation matrix

of the camera extrinsic parameters for the first frame, and K is the camera intrinsic parameters. Then, as is shown in Fig. 3, we plot the ego vehicle trajectory and other vehicle trajectories in separate blank videos to generate trajectory videos. For frame t in the trajectory video, only current and future points $\text{Traj}^{plot} = \{(x_t^I, y_t^I), ..., (x_T^I, y_T^I)\}$ are plotted, indicating the future motion of the vehicle. As a result, the trajectory video contains the motion information of vehicles and corresponds each trajectory to the vehicle in the video.

Aligned Motion Guidance Generation

After obtaining trajectory videos, the trajectory can be learned in unified visual modality with driving scene videos. The world model needs to encode trajectory videos as conditions to guide video generation. Here, we adopt the Spatial-Temporal Variational Auto Encoder (STVAE) in CogvideoX (Yang et al. 2024b) to encode them. On the one hand, STVAE is capable of extracting high-quality spatial-temporal features of trajectory videos, which are aligned with driving video latents temporally and spatially. On the other, the condition features can share the feature space with the driving video latents since they both use the same STVAE to generate latents, which is easier to be learned jointly. The vehicle trajectory latents z_{traj} and original video latents z_{vid} for training can be formulated as follows:

$$z_{traj} = \Psi(V_{traj}), \tag{2}$$

$$z_{vid} = \Psi(V_{ori}), \tag{3}$$

where V_{traj}, V_{ori} are the ego vehicle trajectory video and original driving video, respectively. Ψ denotes the STVAE in the model. z_{traj} is used to provide motion guidance for video generation.

Trajectory-injected Diffusion Transformer

Typical Diffusion Transformer in CogvideoX (Yang et al. 2024b) aims to predict the noise n^a added to z_{vid} based on the noisy video latent z_{vid}^{noise}, which is formulated as

$$z_{vid}^{noise} = z_{vid} + n^a. \tag{4}$$

In the inference stage, Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020) can denoise from random sampling noise to video latents for the VAE decoder to generate video. However, the video cannot be generated with specific initial frames in such a manner, which is important for the world model. In addition, the original CogvideoX (Yang et al. 2024b) can only use the text as the condition to control video generation.

Therefore, we design Trajectory-injected Diffusion Transformer (TiDiT) to predict future driving scenes based on specific initial frames and trajectories. Specifically, to enable the model to generate the video from historical T_c frames, we replace the first T_c frames in noisy video latents z_{vid}^{noise} with the first T_c frames of z_{vid}. The process can be realized via a conditioning mask $M = \{t \leq T_c | 1 \leq t \leq T\}$, which is formulated as

$$z = M * z_{vid} + (1 - M) * z_{vid}^{noise}. \tag{5}$$

To further inject trajectory conditions, frame latents z and motion guidance z_{traj} sharing the feature space are concatenated in channel dimension. As a result, the video and condition latents are fused and aligned spatially and temporally. Then, 3D convolution layers are adopted for learning to obtain the final visual latents z_{vid}^f, which can be calculated as

$$z_{vid}^f = \text{3d\_conv}(\text{concat}([z, z_{traj}])). \tag{6}$$

The original Diffusion Transformer in CogvideoX (Yang et al. 2024b) uses the text as the condition to control video generation. To fully reserve the ability of CogvideoX, we utilize the scene caption provided in OmniDrive (Wang et al. 2024a) to train the proposed TiDiT. As a result, visual latents z_{vid}^f and text embedding z_{text} extracted from the scene caption are concatenated and fed into several layers of Expert Diffusion Transformer in CogvideoX to predict the noise added to z_{vid} as follows:

$$n^p = \phi(\text{concat}([z_{vid}^f, z_{text}])), \tag{7}$$

where ϕ represents the Expert Diffusion Transformer in CogvideoX. Given the added noise n^a and the predicted noise n^p, the diffusion loss l of the proposed model can be formulated as

$$l_{diff} = \frac{\sum_{i=1}^{T}(n_i^p - n_i^a) * M_i}{T}. \tag{8}$$

In the inference stage, we use the STVAE to extract latents of the initial frames, and replace the corresponding frames of the random sampling noise to obtain z. Then, following Eq. 6 and Eq. 7, the Expert Diffusion Transformer infers the noise for the VAE decoder to generate video.

Evaluate Trajectory Controllability in Video Latent Space. To quantitatively and qualitatively evaluate trajectory controllability, we propose a metric based on control latent similarity in the video latent space. Specifically, we first calculate the predicted video latents z_{vid}^f according to the prediction of TiDiT, which can be formulated as

$$z_{vid}^p = z_{vid}^{noise} - n^p. \tag{9}$$

Then, we adopt a two-layer MLP to decode the predicted video latents as the predicted trajectory video latents z_{traj}^p as follows

$$z_{traj}^p = \Omega(z_{vid}^p), \tag{10}$$

where Ω represents the MLP. We design a control latent similarity loss l_{traj} to train the MLP, which is formulated as

$$l_{traj} = \frac{\sum_{i=1}^{T}(z_{traj}^p - z_{traj})}{T}, \tag{11}$$

which also serves as the metric to evaluate trajectory controllability. Moreover, the predicted trajectory video latents z_{traj}^p can be decoded via STVAE to generate trajectory video for visualization.

Experiments

Dataset and Evaluation Metric

Extensive experiments are conducted on the nuScenes dataset (Caesar et al. 2020) to evaluate the proposed EOT-WM. Following the setup of Vista (Gao et al. 2024) in the

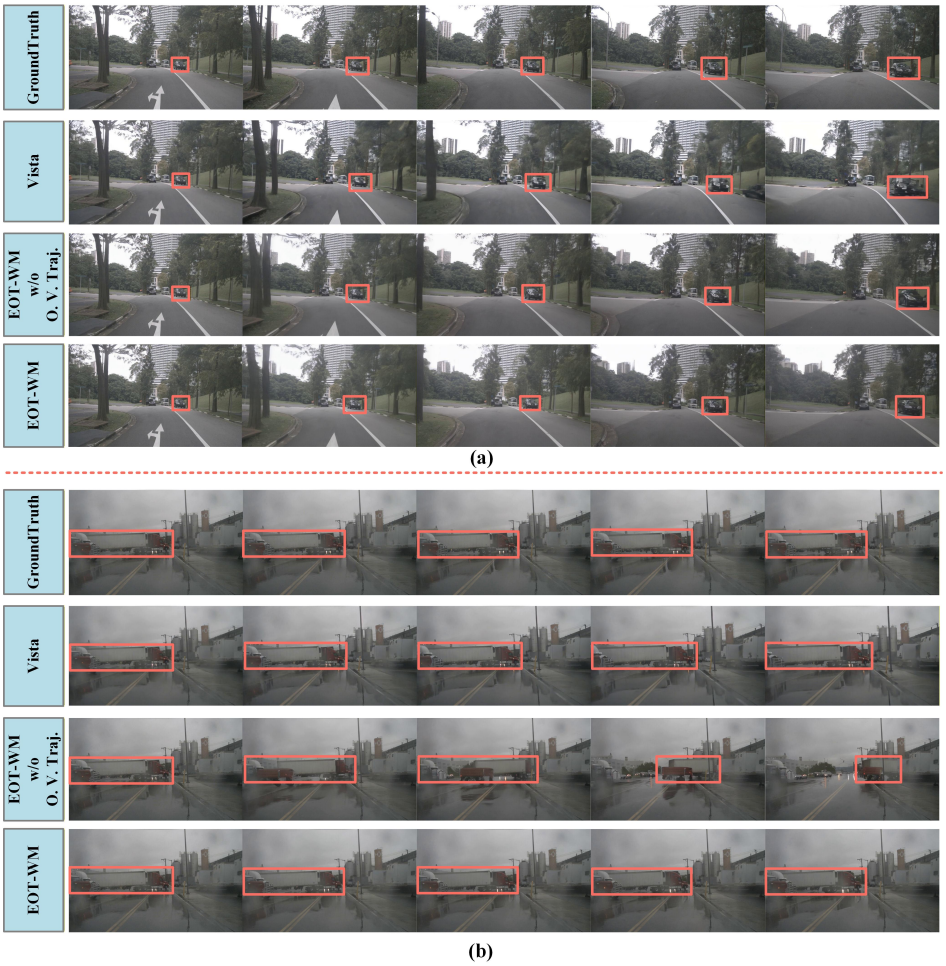

Figure 4: Representative cases for action controllability achieved by the proposed EOT-WM and Vista on the validation set of nuScenes dataset, where EOT-WM w/o O.V. Traj. means the proposed model without learning other vehicle trajectories.

nuScenes dataset, the training and evaluation set contain 25109 and 5369 videos with 25 frames. The ego vehicle trajectory corresponding to the video can be directly obtained in the annotations. The other vehicle trajectory is only annotated in key frames, and we use interpolation based on the annotated trajectories to complete the full trajectory corresponding to the 25-frame video. We use FID (Heusel et al. 2017) and FVD (Unterthiner et al. 2018) to evaluate the quality of generated videos. Moreover, we provide qualitative demonstrations that underscore the advancements of the proposed EOT-WM.

Implementation Details

We build the proposed EOT-WM based on CogvideoX-2B (Yang et al. 2024b). We use the first video latent as the context for predicting the future frames. The frame rate of the generated video is 10Hz. We use the AdamW (Loshchilov 2017) optimizer with a learning rate of 2×10^{-5} to train the model. By default, the proposed model is trained at 768×1280 resolution for 60 epochs with 64 NVIDIA A800 GPUs with a total batch size of 128.

Comparison with State-of-the-art Methods

Quantitative Result. Comparison results of FID and FVD for video generation achieved by the proposed EOT-WM and competing methods on the validation set of nuScenes dataset are given in Tab. 1. As can be observed, the state-of-the-art driving world model Vista (Gao et al. 2024) can generate videos with higher quality and resolution compared with previous works (Santana and Hotz 2016; Wang et al. 2023b; Lu et al. 2024; Wang et al. 2024b; Yang et al. 2024a). However, Vista (Gao et al. 2024) cannot control other vehicles in the scene, and using trajectories in BEV space is not sufficiently effective. Therefore, the proposed EOT-WM controlling video generation with ego and other vehicle trajectories in video space achieves 4.8 FID and 40.0 FVD, surpassing Vista by 30% in FID and 55% in FVD. Moreover, the proposed EOT-WM can generate videos in 768×1280 resolution, superior to previous world models as well.

Qualitative Result. To further demonstrate the superiority in action controllability of the proposed EOT-WM, we provide several representative cases achieved by the proposed EOT-WM compared with Vista (Gao et al. 2024) in

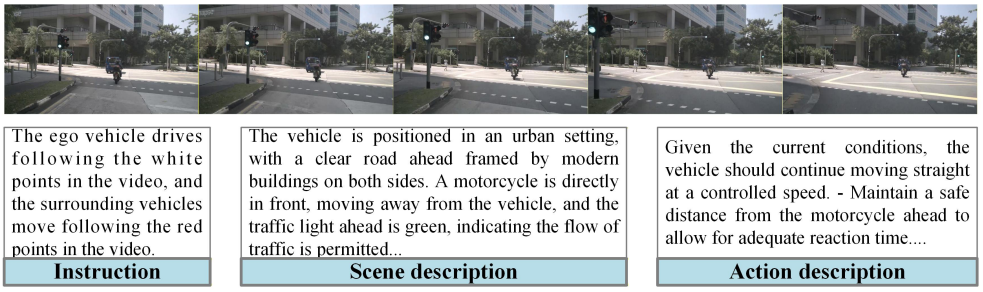

Figure 5: Instance of difference text types used for the proposed EOT-WM, where the scene description and action description are provided in OmniDrive (Wang et al. 2024a).

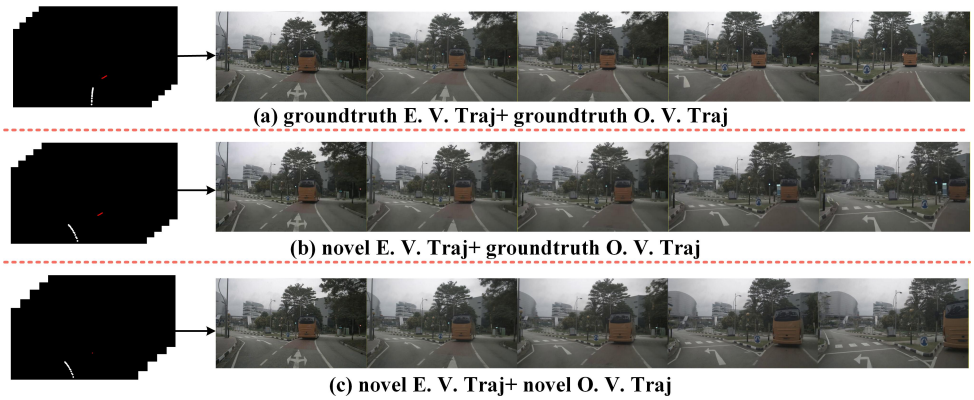

Figure 6: Instances generated with novel trajectories by the proposed EOT-WM.

Methods	Resolution	FID ↓	FVD ↓
DriveGAN (Santana and Hotz 2016)	256×256	73.4	502.3
DriveDreamer (Wang et al. 2023b)	128×192	52.6	452.0
WoVoGen (Lu et al. 2024)	256×448	27.6	417.7
Drive-WM (Wang et al. 2024b)	192×384	15.8	122.7
GenAD (Yang et al. 2024a)	256×448	15.4	184.0
Vista (Gao et al. 2024)	576×1024	6.9	89.4
EOT-WM	768×1280	**4.8**	**40.0**

Table 1: Comparison results of FID and FVD for video generation achieved by the proposed EOT-WM and competing methods on the validation set of nuScenes dataset.

VTR	AMGG	E. V. Traj.	O. V. Traj.	FID ↓	FVD ↓
-	-	-	-	61.4	372.8
-	✓	✓	✓	55.3	348.1
✓	-	✓	✓	26.1	227.2
✓	✓	-	✓	49.6	312.5
✓	✓	✓	-	7.2	66.1
✓	✓	✓	✓	**4.8**	**40.0**

Table 2: Ablation for key components on the validation set.

Fig. 4. As is shown in Fig. 4 (a), the blank car in the red rectangle stands by the side of the road. Vista and EOT-WM w/o O.V. Traj. do not consider the other vehicle trajectories, and the blank car moves forward in the generated videos. In contrast, the blank car keeps still in the video generated by the proposed EOT-WM. In Fig. 4 (b), the truck is backing up on the road. However, Vista and EOT-WM w/o O.V. Traj. both generate the truck moving forward in the generated video. As for the proposed EOT-WM, the action of the truck is predicted correctly in the generated video. These cases demonstrate that the proposed EOT-WM can generate more realistic videos with ego and other vehicle trajectories.

Ablation Study

To verify the effectiveness of the proposed modules, the ablation study for key components is conducted on the validation set of nuScenes dataset, and the results are given in Tab. 2. When all components are not used, the model is close to action-free only with initial frames as the context. As a result, the model obtains poor performance as 61.4 FID and 372.8 FVD. As for **EOT-WM w/o VTR** model, the performance is still poor, since the model cannot learn trajectories in BEV space well and other modules designed for VTR cannot work. For **EOT-WM w/o AMGG** model, we use C3D to extract trajectory video features, and the performance declines compared with the proposed EOT-WM due to the dissimilarity of the feature space. The ego vehicle tra-

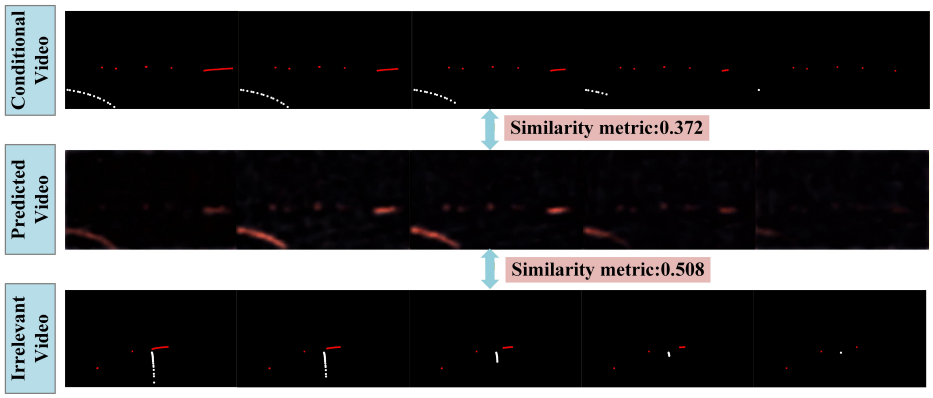

Figure 7: Instance of the proposed metric for evaluating trajectory controllability.

jectory is the most important condition for video generation. Therefore, **EOT-WM w/o E. V. Traj.** model cannot generate realistic videos. With other vehicle trajectories to control the other vehicles in the scene, the proposed **EOT-WM** model can achieve better FID and FVD compared with **EOT-WM w/o O. V. Traj.** model. All above results demonstrate the effectiveness of the proposed modules.

Effect of difference text types

The original CogvideoX (Yang et al. 2024b) is trained to use the text to control video generation, and other advanced video generation models (Guo et al. 2024; Kong et al. 2024) are also text-to-video models. When building the world model based on such text-to-video models, it is unadvisable to remove the text element since the capability of the original text-to-video model will be damaged. As can be observed in Tab. 3, while the proposed EOT-WM does not use the text, FID and FVD achieved decline obviously. Then, we adopt three types of texts to control the video generation model. The performances obtained by these types of texts are similar, demonstrating that the video generation model can be trained to adapt different types of texts. Among them, EOT-WM using the action description achieves the best performance since the caption contains action information. However, considering the capability of generating novel scenes, we choose the scene description for EOT-WM as the action description will conflict with the novel trajectories.

Generating novel scenes

To verify the capability of the proposed EOT-WM to generate novel scenes, several instances generated with novel trajectories are given in Fig. 6. As can be observed in Fig. 6 (a), the ego vehicle turns right and the bus moves forward in the video generated with groundtruth ego and other vehicle trajectories in the dataset. While we use the self-produced ego vehicle trajectory that represents turning left in Fig. 6 (b), the ego vehicle is controlled to turn left in the video generated. Moreover, we give the novel other vehicle trajectory meaning stop in Fig. 6 (c), the bus in the generated video stands by. Above results demonstrate that the proposed EOT-WM can generate novel scenes according to novel trajectories.

Evaluating Trajectory Controllability

To evaluate trajectory controllability, we propose a metric based on control latent similarity in the video latent space. As is shown in Fig. 7, we use the conditional trajectory video to generate predicted trajectory video latents in TiDiT. The predicted trajectory video latents can be decoded into predicted trajectory video, which is similar to the conditional trajectory video. Moreover, given another irrelevant trajectory video, the metric calculated between the predicted trajectory video latents and the conditional trajectory video latents is obviously smaller than that calculated between the predicted trajectory video latents and the irrelevant trajectory video latents. These results demonstrate the effectiveness of the proposed latent-based metric.

Text Type	FID ↓	FVD ↓
None	18.5	156.3
Instruction	5.2	42.1
Scene Description	4.8	40.0
Action Description	**4.1**	**37.6**

Table 3: Effect of difference texts used in EOT-WM.

Conclusion

In this work, a novel driving world model named EOT-WM is proposed, which can use ego and other vehicle trajectories to generate realistic videos. Specifically, we propose to represent ego and other vehicle trajectories in the video space instead of the BEV space for learning in the unified visual modality. Then, AMGG is proposed to generate trajectory latents aligned with driving video latents spatially and temporally, which also share feature space. Finally, TiDiT is designed to denoise the noisy video latents more precisely via jointly using the motion guidances based on the ego-other vehicle trajectories. Experiments conducted on the nuScenes dataset demonstrate the superiority of the proposed EOT-WM. In the future work, we will explore the manner to control the other vehicles more precisely and the quantitative metric to evaluate the consistency between the given trajectory and the generated video.

References

Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; et al. 2025. Cosmos world foundation model platform for physical ai. *in arXiv:2501.03575*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *in arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; et al. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. CVPR'23*, 22563–22575.

Brooks, T.; Hellsten, J.; Aittala, M.; Wang, T.-C.; et al. 2022. Generating long videos of dynamic scenes. In *Proc. NeurIPS'22*, 31769–31781.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; et al. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proc. CVPR'20*, 11621–11631.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *in arXiv:2310.19512*.

Dorkenwald, M.; Milbich, T.; Blattmann, A.; Rombach, R.; et al. 2021. Stochastic image-to-video synthesis using cinns. In *Proc. CVPR'21*, 3742–3753.

Franceschi, J.-Y.; Delasalles, E.; Chen, M.; Lamprier, S.; and Gallinari, P. 2020. Stochastic latent residual video prediction. In *Proc. ICML'20*, 3233–3246.

Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; et al. 2024. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. In *Proc. NeurIPS'24*.

Ge, S.; Hayes, T.; Yang, H.; Yin, X.; et al. 2022. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *Proc. ECCV'22*, 102–118.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; et al. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *Proc. ICLR'24*.

Ha, D.; and Schmidhuber, J. 2018. Recurrent world models facilitate policy evolution. In *Proc. NeurIPS'18*.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *Proc. ICLR'20*.

Hafner, D.; Lillicrap, T. P.; Norouzi, M.; and Ba, J. 2021. Mastering Atari with Discrete World Models. In *Proc. ICLR'21*.

Hassan, M.; Stapf, S.; Rahimi, A.; Rezende, P.; et al. 2025. GEM: A Generalizable Ego-Vision Multimodal World Model for Fine-Grained Ego-Motion, Object Dynamics, and Scene Composition Control. In *Proc. CVPR'25*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. NeurIPS'17*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proc. NeurIPS'20*, 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; et al. 2022. Video diffusion models. In *Proc. NeurIPS'22*, 8633–8646.

Hu, A.; Murez, Z.; Mohan, N.; Dudas, S.; et al. 2021. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proc. CVPR'21*, 15273–15282.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; et al. 2023. Planning-oriented autonomous driving. In *Proc. CVPR'23*, 17853–17862.

Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; et al. 2023. Vad: Vectorized scene representation for efficient autonomous driving. In *Proc. ICCV'23*, 8340–8350.

Koh, J. Y.; Lee, H.; Yang, Y.; Baldridge, J.; and Anderson, P. 2021. Pathdreamer: A world model for indoor navigation. In *Proc. ICCV'21*, 14738–14748.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; et al. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *in arXiv:2412.03603*.

Kumar, M.; Babaeizadeh, M.; Erhan, D.; Finn, C.; et al. 2020. VideoFlow: A Conditional Flow-Based Model for Stochastic Video Generation. In *Proc. ICLR'20*.

Li, H.; Sima, C.; Dai, J.; Wang, W.; et al. 2024. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE Trans. Pattern Anal. Mach. Intell*, 46(4): 2151–2170.

Li, T.; Wang, H.; Li, X.; Liao, W.; et al. 2025. Generative Planning with 3D-vision Language Pre-training for End-to-End Autonomous Driving. *in arXiv:2501.08861*.

Loshchilov, I. 2017. Decoupled weight decay regularization. *in arXiv:1711.05101*.

Lu, J.; Huang, Z.; Yang, Z.; Zhang, J.; and Zhang, L. 2024. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *Proc. ECCV'24*, 329–345.

Mendonca, R.; Bahl, S.; and Pathak, D. 2023. Structured world models from human videos. *in arXiv:2308.10901*.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; et al. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proc. ICML'22*, 16784–16804.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR'22*, 10684–10695.

Santana, E.; and Hotz, G. 2016. Learning a driving simulator. *in arXiv:1608.01230*.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; et al. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *in arXiv:2402.12289*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; et al. 2018. Towards accurate generative models of video: A new metric & challenges. *in arXiv:1812.01717*.

Villegas, R.; Pathak, A.; Kannan, H.; Erhan, D.; et al. 2019. High fidelity video prediction with large stochastic recurrent neural networks. In *Proc. NeurIPS'19*.

Wang, H.; Liang, W.; Van Gool, L.; and Wang, W. 2023a. Dreamwalker: Mental planning for continuous vision-language navigation. In *Proc. ICCV'23*, 10873–10883.

Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; et al. 2024a. OmniDrive: A Holistic LLM-Agent Framework for Autonomous Driving with 3D Perception, Reasoning and Planning. *in arXiv:2405.01533*.

Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; et al. 2023b. Drivedreamer: Towards real-world-driven world models for autonomous driving. *in arXiv:2309.09777*.

Wang, Y.; He, J.; Fan, L.; Li, H.; et al. 2024b. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proc. CVPR'24*, 14749–14759.

Wang, Z.; Yuan, Z.; Wang, X.; Li, Y.; et al. 2024c. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH'24*, 1–11.

Weissenborn, D.; Täckström, O.; and Uszkoreit, J. 2020. Scaling Autoregressive Video Models. In *Proc. ICLR'20*.

Yang, J.; Gao, S.; Qiu, Y.; Chen, L.; et al. 2024a. Generalized predictive model for autonomous driving. In *Proc. CVPR'24*, 14662–14672.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; et al. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *in arXiv:2408.06072*.

Yu, S.; Tack, J.; Mo, S.; Kim, H.; et al. 2022. GENERATING VIDEOS WITH DYNAMICS-AWARE IMPLICIT GENERATIVE ADVERSARIAL NETWORKS. In *Proc. ICLR'22*.

Zhang, Z.; Liao, J.; Li, M.; Dai, Z.; et al. 2024. Tora: Trajectory-oriented diffusion transformer for video generation. *in arXiv:2407.21705*.