Fine-flow Distilling Coarse-flow Video Generation for Long-Term Driving World Model

Xiaodong Wang[1,2], **Zhirong Wu**[1], **Peixi Peng**[1,2*]
[1]School of Electronic and Computer Engineering, Peking University
[2]Pengcheng Laboratory
{wangxiaodong21s@stu., pxpeng@}pku.edu.cn

Abstract

Driving world models are used to simulate futures by video generation based on the condition of the current state and actions. However, current models often suffer serious error accumulations when predicting the long-term future, which limits practical applications. Recent studies utilize the Diffusion Transformer (DiT) as the backbone of driving world models to improve learning flexibility. However, these models are always trained on short video clips, and multiple roll-out generations struggle to produce consistent and reasonable long videos due to the training-inference gap. To this end, we propose several solutions to build a simple yet effective long-term driving world model. First, we hierarchically decouple world model learning into large motion learning and bidirectional continuous motion learning. Then, considering the continuity of driving scenes, we propose a simple distillation method where fine-grained video flows are self-supervised signals for coarse-grained flows. The distillation is designed to improve the coherence of infinite video generation. The coarse-grained and fine-grained modules are coordinated to generate long-term and temporally coherent videos. On NuScenes, compared with the state-of-the-art front-view models, our model improves FVD by 27% and reduces inference time by 85% for the video task of generating 110+ frames.

Code —
https://github.com/Wang-Xiaodong1899/Long-DWM

Introduction

World models are used to predict the environment dynamics of different actions based on the current state (Ha and Schmidhuber 2018; LeCun 2022), which is very important for autonomous driving. Earlier works design world models in latent feature space (Hafner et al. 2019, 2020, 2023), and could facilitate the learning of control policies (Ebert et al. 2018; Dosovitskiy et al. 2017; Tassa et al. 2018). To improve the interpretability and interoperability to the human driver, several works use controllable video generation technology to build driving world models (Hu et al. 2023; Wang et al. 2023b,c; Gao et al. 2024b), and video generators also achieved promising results as world simulators (OpenAI 2024; Kong et al. 2024). For driving world models, long-term prediction capability is essential to deliver accurate and reliable guidance for current decision-making.

However, serious error accumulations are easy to observe when driving world models predict the long-term future, which limits the practical application. The essential reason is that long video generation is a challenging task, and needs to bridge the gap between general scenarios and driving scenarios, especially in driving tasks with large motion. It requires generated videos to have long-term coherent, reasonable and accurate scene development. The first challenge comes from the training-inference gap. Current driving world models train diffusion models in short clips with high fps, such as Vista (Gao et al. 2024b), a state-of-the-art model based on a U-Net backbone trained on 25-frame clips with 10 fps, which encounters serious error accumulation when rolling out long videos, as shown in the first row in Fig. 1. Training with short clips and continuous frames can be regarded as a "free lunch", since frames with short duration and high fps with smoother temporal distribution for easier learning, and it is widely used in driving or the general domain, such as a state-of-the-art model CogVideoX (Yang et al. 2024b) based the Diffusion Transformer (DiT) backbone (Peebles and Xie 2023). But these models tend to generate only smoother temporal motion and often produce significant error accumulation.

The second challenge is the gap between general scenarios and driving scenarios. If we use a general video generator such as CogVideoX to roll out long videos from a static driving scene, as shown in the second row in Fig. 1, although it is a powerful model utilizing DiT in open-domain, and alleviates the accumulation of edge distortion to some extent, it still produces blurred frames and unrealistic motion. The potential reason is that general scenarios and driving scenarios differ significantly in some aspects, such as motion and scene development. A direct way to bridge the gap is finetuning in the driving scenario, such as adapting CogVideoX (Yang et al. 2024b) into CogVideoX$^{\texttt{sft}}$. However, unrealistic motion artifacts persist in this model, as shown in the third row in Fig. 1. These two challenges have not yet been addressed. This study attempts to address these challenges and builds a simple long-term driving world model, and our model's long video prediction is shown in the last row in Fig. 1. Our prediction result is significantly better than others, showing clear details and consistent temporal dynamics.

*Corresponding author.

Figure 1: Long video generation comparison in the autonomous driving scenario. The results include Vista (Gao et al. 2024b), CogVideoX (Yang et al. 2024b), CogVideoX$^{\texttt{sft}}$ and ours, the input first frame (f=1) is from the validation set of NuScenes (Caesar et al. 2020). The result of Vista shows extremely blurred frames (e.g., yellow stripe), and the results of CogVideoX and variant show unrealistic motion, while our model generates a realistic long-term future.

To address the above challenges, we propose several solutions. For failure patterns of training-inference gap, previous works (Zhao et al. 2025) point out that unrealistic motion exist when rolling out long videos using CogVideoX and HunyuanVideo (Kong et al. 2024). This phenomenon also exists in the predictions of driving world models, such as the error accumulations in the results of Vista and CogvideoX$^{\texttt{sft}}$. To eliminate the gap, some related works (Zhao et al. 2024; Yin et al. 2023) use textual scripts to generate keyframes for a long video and use image-to-video models to extend keyframes. But these works only generate movies or cartoons. To address the problem for driving world models, we decouple the long-term learning into large motion learning (e.g., scene transitions) and small continuous motion learning (e.g., car motions). Decoupled learning naturally eliminates the gap, because world models can first predict large motions and then fill in with small continuous motions for driving.

For the gap between general scenarios and driving scenarios, previous works (Zhao et al. 2024; Yin et al. 2023) only focus on text-to-video, emphasizing that the plot development of long videos needs to conform to given scripts. While in driving world models, the target is to predict the long-term future from the current state, i.e., the current image, and the results of large motion prediction should maintain coherence with the current state. To achieve this, for large motion learning, we construct a Coarse DiT training on coarse frames for causal prediction, and video tokens are encoded independently from each video frame using a low fps. No temporal compression is used to preserve the details of each frame. For small continuous motion learning, a Fine DiT is trained on temporally compressed video tokens for causal and bidirectional predictions, where the video frames have a high fps and adjacent frames are compressed. Meanwhile, considering the continuity characteristics of driving scenes, simply predicting large motions leads to scene mutations and distortion. To this end, we propose a simple flow distillation method. Intuitively, fine-grained video tokens have better consistency than coarse-grained video tokens, so we use the fine-grained priors to guide the coarse-grained prediction. Given well-trained Fine DiT and Coarse DiT, we first sample a continuous frame sequence where coarse frames are marked. We input all coarse frames into a trainable Coarse DiT to obtain the coarse flow, and input all continuous frame segments where the first and last are coarse frames into the frozen Fine DiT to obtain fine flows, where the flow is a one-step denoising output. The distillation loss is an L2 loss between coarse flows and fine flows at corresponding positions. This distillation regards fine flows as self-supervised signals, thereby encouraging the Coarse DiT to make more consistent predictions across frames. Additionally, we propose a novel warp-guided controllable video prediction method for Fine DiT to improve the controllability. Our contributions are three-fold:

- proposes a simple yet effective long-term driving world model, where the world model learning is firstly decoupled into hierarchical learning, including large motion learning and bidirectional continuous motion learning.
- proposes a flow distillation method that fine video flows are self-supervised signals for coarse video flows, prompting the consistency of coarse token predictions.
- Experiments on NuScenes dataset demonstrate our model achieves state-of-the-art performances on the FVD metric in all settings. In particular, compared with the state-of-the-art front-view model Vista, our model improves FVD by 27% and reduces inference time by 85% for the video task of generating 110+ frames.

Related Work

Video Generation

Video Diffusion Models Diffusion models have made great progress in video generation. Early works usually leverage a pretrained text-to-image model (Rombach et al. 2022) and insert temporal layers into the base architecture and continue train on video-text paired data (Wang et al. 2023a; Blattmann et al. 2023b; Guo et al. 2023; Blattmann et al. 2023a). (Blattmann et al. 2023b) inserts temporal convolution and temporal layers into base diffusion U-Net to adapt the video generation task. (Yin et al. 2023) also inserts various temporal layers and other conditional layers into the base architecture. Some works only train extra temporal layers (Blattmann et al. 2023b; Guo et al. 2023) and some works fine-tune the full models (Blattmann et al. 2023a). Recently, diffusion transformer (DiT) models have shown great improvement in video quality. (Peebles and Xie 2023) utilizes Transformer as the backbone of diffusion models, which prompts the text-to-video to reach a new milestone such as Sora (OpenAI 2024). Following Sora, there are some impressive open-sourced models, such as Vidu (Bao et al. 2024), CogVideoX (Yang et al. 2024b), HunyuanVideo (Kong et al. 2024), etc. These work try to bridge the gap between the open-source models and closed-source models.

Long Video Generation The main challenge for this task is that long videos encounter error accumulation, resulting in blurring and distortion of videos. Naturally, auto-regressive models are more suitable for this since they can receive variable video context and generate video frames with variable length, which can alleviate error accumulation using slide windows (Liang et al. 2022; Henschel et al. 2024), but also face high memory pressure for a long sequence of video tokens. Diffusion models are always trained on video frames with a fixed length, and most works train their models on clips with high fps and short duration, due to GPU memory limitation (Yang et al. 2024b; Kong et al. 2024; Bao et al. 2024). Some works utilize hierarchical approaches that first generate keyframes and then interpolate continuous frames between them (Yin et al. 2023; Zhao et al. 2024). However, these works only focus on text-to-video, emphasizing that the plot development of long videos needs to conform to given scripts. For driving world models, future predictions should maintain coherence with the current state. We propose a novel distillation method to guide the long video generation, instead of treating different granularities independently.

Driving World Model

Driving world models leverage the world model to predict the environment dynamics of different driving actions based on the current state. As predictive and generative models have achieved great progress, based on these, driving world models have better instruction-following capabilities and higher-quality predictions. GAIA-1 (Hu et al. 2023) proposes a driving world model based on diffusion models that leverages video, text, and action inputs to generate future scenarios. Some works (Wang et al. 2023b; Jia et al. 2023; Wang and Peng 2025) build multi-modal driving world models to support video generation and action prediction. Besides video and text, some works (Wang et al. 2023c; Gao et al. 2023) utilize 3D annotations or multi-view inputs to predict future scenarios with nuanced 3D geometry. Recent works expand driving world models to the evolution predictions of 3D occupancy (Zheng et al. 2025; Xu et al. 2025) or holistic models for perception, prediction and planning (Zheng et al. 2024). Vista (Gao et al. 2024b) proposes a driving world model with higher resolution and versatile controllability that can generalize to diverse scenarios. The most related works (Wu et al. 2024; Gao et al. 2024a; Jiang et al. 2024; Guo et al. 2025; Li et al. 2024) focus on multi-view generation or occupancy prediction using DiT, but train with short clips. However, these models face error accumulations when predicting the long future, while our method systematically addresses this.

Methodology

Coarse Diffusion Transformer

Previous methods train diffusion U-Nets or Transformers on video clips with short duration and high fps, since these types of clips have smoother temporal distribution for learning, but these models are faced with error accumulation after multiple rollouts for long video generation (Gao et al. 2024b). To avoid relying on the above "free lunch" approach, this paper proposes a novel Coarse Diffusion Transformer (CDiT) trained on a few frames from long-duration clips. The CDiT model offers more robust forward-looking predictions and is capable of predicting large dynamic information.

Given a long video clip v, we first sample K frames using a low fps (such as fps=1), denoted as coarse video frames $v^c = \{v^{c1}, v^{c2}, \cdots, v^{cK}\}$. There is greater dynamic information between these coarse-grained frames, which the CDiT model needs to learn. These coarse video frames are no longer suitable for causal encoding by 3D-VAE, so we use 3D-VAE to encode each coarse frame independently. The coarse video latents are denoted as $x^c = \{x^{c1}, x^{c2}, \cdots, x^{cK}\}$. Given the latent x_0 drawn from data distribution $q(x_0)$, and $x_1, \cdots, x_T$ are latents of the same dimensionality as x_0, the diffusion process is defined as a Markov chain as below:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where $\alpha_t > 0$ is a scalar according to a specific noise scheduler at timestep t. We regard $x_0^c := x^c$ as the start point of the forward process, and add random Gaussian noise ϵ:

$$x_t^c = \sqrt{\bar{\alpha}_t}x_0^c + (1-\bar{\alpha}_t)\epsilon,\ \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where t is a random timestep and $\bar{\alpha}_t := \Pi_{s=1}^{t}\alpha_s$.

The large dynamic information requires more fine-grained video descriptions. Unlike Vista (Gao et al. 2024b) that uses overly simplistic video annotations, we utilize a multi-modal large language model (MLLM) (Zhang et al. 2024) to annotate videos with more detailed annotations. More details can be found in the Appendix. Given a detailed video prompt, we utilize T5 (Raffel et al. 2020) to encode the prompt to a text embedding p. Then, we design a image-to-video prediction task in the latent space as follows:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x^c,p,\epsilon,t}\|D_\theta^c(x_t^c, t, p, x^{c1}) - x^c\|_2^2, \tag{3}$$

where D_θ^c is the CDiT denoiser that shares the same architecture with CogVideoX (Yang et al. 2024b), and x^{c1} denotes

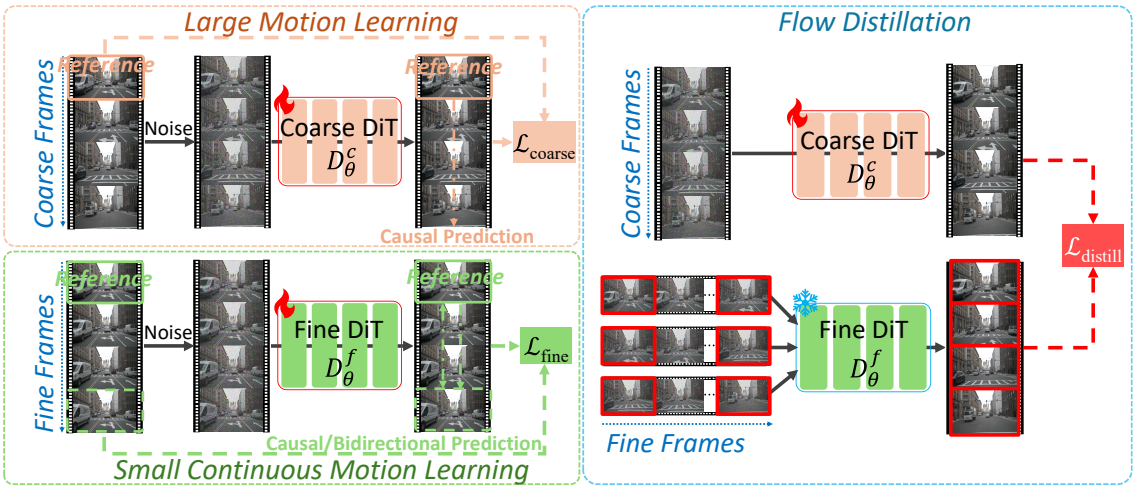

Figure 2: The overall distillation framework. First, we decouple the long-term world model learning into a large motion learning and a small continuous motion learning by designing a Coarse DiT and a Fine DiT to adapt to different granularities. Then, we propose a novel flow distillation method between different granularities, i.e., using fine flows' better priors to distill coarse flows, which prompts the Coarse DiT to produce more consistent predictions. After a few hundred steps, our distillation process effectively tunes the Coarse DiT well.

the first coarse frame of a random sampling clip, and the embedded way is the same as (Wang et al. 2023a).

Learning the large dynamic information would affect the structure and details of each individual frame, so we introduce a latent structure preservation loss as follows:

$$\mathcal{L}_{\text{struct}} = \mathbb{E}_{x^c,p,\epsilon,t}\|\mathcal{H}(D_\theta(x_t^c,t,p,x^{c1})) - \mathcal{H}(x^c)\|_2^2, \quad (4)$$

where $\mathcal{H}$ denotes the 2D high-pass filter of frequency domain in latent space. The final learning objective of CDiT is defined as below:

$$\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{diffusion}} + \beta_s \mathcal{L}_{\text{struct}}, \quad (5)$$

where β_s controls the importance of the preservation.

Fine Diffusion Transformer

Besides the ability to predict large dynamic information, fine-grained details and reasoning information should also be included in the long video prediction process. For example, if there is a car around the ego car in the previous keyframe, but the car is gone in the current keyframe, the correct reasoning should be to fill in a segment of the car slowly driving away between the two frames. To learn the video reasoning ability, we design a versatile Fine DiT (FDiT) that learns both video prediction and video interpolation. In this stage, we sample K continuous frames from short-duration video clips, denoted as $v^f = \{v^{f1}, v^{f2}, \cdots, v^{fK}\}$. The continuous frames are encoded to latents $x^f = \{x^{f1}, x^{f2}, \cdots, x^{fK}\}$ by a 3D-VAE. We reuse the MLLM's annotations and design the training objective as below:

$$\mathcal{L}_{\text{fine}} = \mathbb{E}_{x^f,p,\epsilon,t}\|D_\theta^f(x_t^f,t,p,x^{f1},\ddot{x}^{fK}) - x^f\|_2^2, \quad (6)$$

where D_θ^f denotes the FDiT denoiser, $\ddot{x}^{fK}$ denotes the last frame x^{fK} that would be dropped with a fixed ratio, and FDiT learns causal prediction and bidirectional prediction, that is, video prediction and video interpolation, and the two tasks can help each other.

Fine-flow → Coarse-flow Distillation

The CDiT would sacrifice consistency to achieve large dynamic prediction, unlike previous hierarchical U-Net models neglecting this problem (Yin et al. 2023), we instead propose a novel distillation method that distills prior knowledge from fine flow into the coarse flow of CDiT. Previous methods utilize distillation to reduce denoising steps or computing time in the same granularity flows, i.e., short-duration videos (Yin et al. 2024b,a,c). We propose a simple distillation method between different granularity. The fine flows in FDiT can give more guidance to coarse flows since FDiT is good at predicting fine-grained smooth temporal changes.

The overall illustration is shown in Fig 2. After training both CDiT and FDiT models, we can initialize the distillation process. Given a video clip v, we first sample a sequence of frames using the same high fps as FDiT, and coarse frames are marked, the sequence is denoted as below:

$$v_d = \{\underbrace{v^{c1}, v^{f2}, \cdots, v^{c2}}_{K \text{ items}}, \cdots, v^{cK}\}, \quad (7)$$

where v_d can be divided into segments of $K-1$ consecutive frames, and the first and last frames of each segment are coarse frames. To distill the CDiT to have more consistent predictions, we utilize the frozen FDiT to separately cope with latents of each segment, and then the first and last latent of each segment are concatenated together. Specifically, we first randomly sample a timestep t and a Gaussian noise with the same shape of latents. Then, each segment share the same timestep and noise to predict the latents using one-step denoising. The coarse frames $\{v^{c1}, v^{c2}, \cdots, v^{cK}\}$ are also added by same noise with timestep t and predict the coarse

Model	Extra data	Backbone	FID↓	FVD↓
DriveGAN (Kim et al. 2021)	✗	GAN	73.4	502.3
DriveDreamer (Wang et al. 2023b)	✗	U-Net	52.6	452.0
WoVoGen (Lu et al. 2025)	✗	U-Net	27.6	417.7
Drive-WM (Wang et al. 2023c)	✗	U-Net	15.8	122.7
GenAD (Yang et al. 2024a)	✗	U-Net	15.4	244.0
GenAD (Yang et al. 2024a)	✓	U-Net	15.4	184.0
Vista (Gao et al. 2024b)†	✓	U-Net	**7.6**	128.5
UniMLVG (Chen et al. 2024)	✓	DiT	30.5	149.7
CogVideoX$^{\texttt{sft}}$ (Yang et al. 2024b)	✗	DiT	15.8	117.0
Ours	✗	DiT	12.3	**102.9**
Reconstruction	✗	DiT	4.9	31.3

Table 1: Comparison of prediction fidelity on NuScenes validation set. Our model outperforms the state-of-the-art driving world models. † denotes our evaluation using open source checkpoints from (Gao et al. 2024b). Reconstruction denotes using VAE from CogVideoX (Yang et al. 2024b).

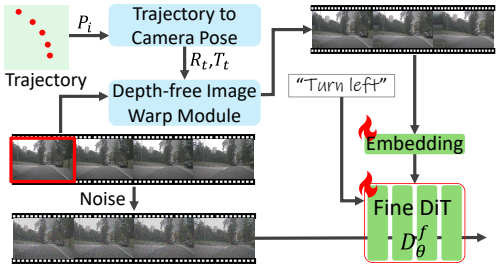

Figure 3: Warp-guided trajectory controllable training. This enhances counterfactual reasoning ability for short videos.

latents. The distillation loss is defined as below:

$$\mathcal{L}_{\text{distill}} = \left\| D^f_{\theta,\text{frozen}}(v_{(d,t)})[c_1, c_2, \cdots, c_K] - D^c_\theta(v_{(d,t)}[c_1, c_2, \cdots, c_K]) \right\|^2 \quad (8)$$

where the teacher model is a frozen FDiT well-trained from Sec. and the student model is initialized from a CDiT well-trained from Sec. . Our distillation requires only a few hundred steps to tune the student model well.

Warp-guided Controllable Video Prediction

Previous methods primarily adapt camera trajectory controllable video generation by fusing trajectory or camera pose features into models, but often perform poorly for complex trajectories (He et al. 2024; Bahmani et al. 2024; Gao et al. 2024b). Recently, some studies (Hou et al. 2024; Bian et al. 2025) improve the generation by incorporating 3D priors. Inspired by this, we propose a novel trajectory control method that leverages 3D information without the need for reconstruction or additional annotations. Given an input trajectory, Vista (Gao et al. 2024b) only inputs it into diffusion models as an extra condition, while in our process, we leverage trajectory priors to obtain warped future frames to stabilize the control ability. More details can be found in Appendix.

Video prediction with warped images. As shown in Fig. 3, our model is built on our Fine DiT and enhances the trajectory controllable video prediction by incorporating warped subsequent frames for guidance. However, the warped frames have some distortions, such as inconsistent rotation and varying motion speeds. These issues stem primarily from our simplistic assumption of uniform depth. To address this, after the warped images are encoded by the 3D-VAE, we first pass them through a trainable patch embedding layer before injecting them into each block of the DiT. Warped features prompt our model to predict more accurate controllable predictions.

Experiments

Implementation Details All training and evaluations are based on NuScenes benchmark (Caesar et al. 2020). Our models are initialized with CogVideoX-2B (Yang et al. 2024b). Then our models are trained with a resolution of 720×480 on the training set. We evaluate the video prediction quality on the validation set utilizing metrics FVD and FID. We choose state-of-the-art models Vista (Gao et al. 2024b), SVD (Blattmann et al. 2023a), DynamiCrafer (Xing et al. 2025), I2VGen-XL (Zhang et al. 2023), and CogVideoX-I2V-5B (Yang et al. 2024b) as baselines for fair comparison. For driving models using DiT, we compare UniMLVG (Chen et al. 2024), since other models (Gao et al. 2024a; Jiang et al. 2024; Guo et al. 2025; Li et al. 2024) only support multi-view predictions. More details refer to the supplementary material.

Results of Generation Quality and Fidelity

Automatic Evaluation We conduct the comparison on all samples of the validation set of NuScenes. Tab. 1 presents the comparison of prediction fidelity of the state-of-the-art models. Compared with previous models without using extra driving videos for training, our model outperforms methods using U-Net or DiT such as GenAD and CogVideoX$^{\texttt{sft}}$ on video fidelity by large gains, and obtains better image fidelity than them. Compared with SOTA models (GenAD, Vista, UniMLVG) using extra driving videos for training, our model also outperforms them on video fidelity. We also report the performance of the reconstruction using CogVideoX VAE, which can be regarded as the FID and FVD lower bounds. The evaluation results show that our models' video prediction

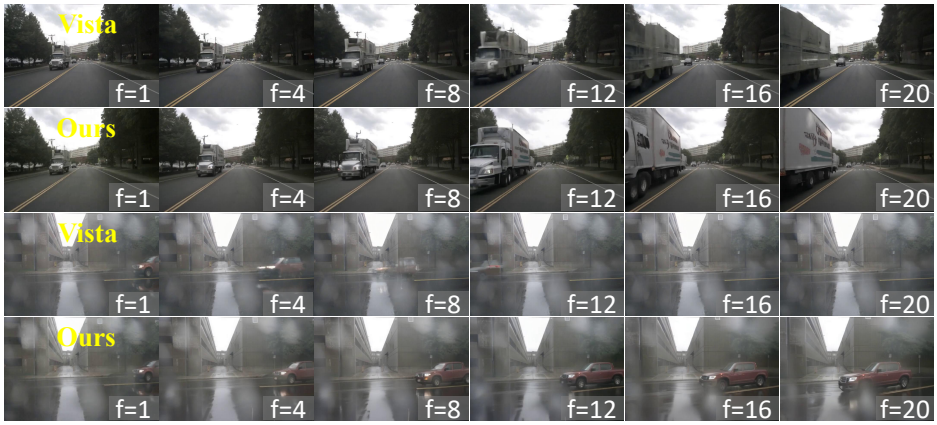

Figure 4: Short video prediction comparison. Compared to our model's prediction with Vista, our model can produce more detailed future frames and generate reasonable and realistic motion.

Duration	**Model**	I2VGen-XL	DynamiCrafter	CogVideoX-I2V-5B	SVD	Vista	Ours
25f∼2.5s(	**FVD**↓	768.1	357.5	400.3	194.6	**174.1**	188.5
1st rollout)	**Time**↓	78s	88s	140s	44s	95s	**35s**
69f∼6.9s(	**FVD**↓	1214.2	602.4	482.5	352.4	277.5	**242.2**
3rd rollout)	**Time**↓	234s	220s	660s	132s	285s	**70s**
113f∼11.3 s(	**FVD**↓	1616.3	1155.5	533.7	753.0	398.5	**289.5**
5th rollout)	**Time**↓	390s	352s	990s	220s	475s	**70s**

Table 2: Long-term video prediction on NuScenes validation subset. We report the FVD and inference time for various durations.

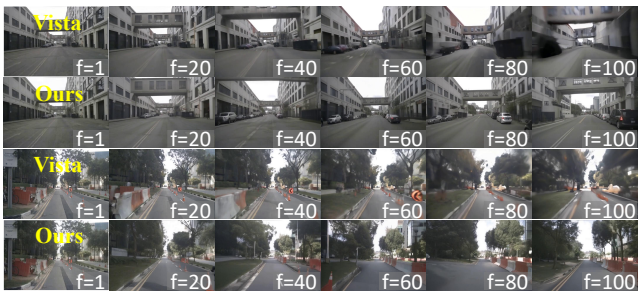

Figure 5: Long-term video prediction comparison. Compared with the state-of-the-art Vista (Gao et al. 2024b), which faces severe error accumulation, our model can predict higher-quality long-term future and generalize to diverse scenarios.

Figure 6: Trajectory controllability comparison. For different input images and trajectory points, we present the generation results (the 12th and 24th frames) of Vista and our model. This highlights the counterfactual reasoning capability.

quality is closer to FVD's lower bounds. The short video prediction comparison is shown in Fig. 4.

Human Evaluation Using automatic metrics does not always align with human preference, so we introduce the human evaluation to asses the visual quality and motion rationality of generated videos following (Gao et al. 2024b). Because Vista significantly outperforms other baselines, based on the human evaluation results in (Gao et al. 2024b), in order to save the cost of manual annotation, we only conducted a human evaluation of random side-by-side video selection between our method and Vista. Details can be found in Appendix. As shown in Tab. 3a, our method outperforms Vista on visual quality and motion rationality.

Results of Long-term Prediction To compare the ability of long-term prediction, we set up comparisons with various video frames including 25, 69, and 113 frames, which correspond to the 1st, 3rd, and 5th rollouts of Vista, respectively. Table. 2 reports FVD and inference time of baselines and our model. As for inference time, other baselines only support autoregressive rollouts, so the time increases linearly, while our model supports parallel interpolation, which reduces the total inference time significantly. Our model achieves comparable FVD with Vista for short videos, but surpasses all baselines on longer durations. The results indicate our model effectively bridges the gap between general and driving scenes.

Model	V.Q.↑	M.R.↑
Vista	45.5%	47.5%
Ours	**54.5%**	**52.5%**

(a)

Model	Win Rate↑
Vista	32.1%
Ours	**67.9%**

(b)

Table 3: Evaluation for (a) visual quality and motion rationality and (b) trajectory compliance.

Model	FID↓	FVD↓
Vista	**9.3**	118.8
CogVideoX + Vista's feature	13.0	89.8
Ours	12.7	**69.6**

Table 4: Trajectory-based evaluation. Trajectory annotation comes from Vista (Gao et al. 2024b).

Results of Controllable Prediction

Automatic Evaluation We set up a fair trajectory-based prediction comparison using Vista's annotations and compare our model with Vista. The results are shown in Tab. 4. Our model outperforms Vista on video quality by a significant margin according to FVD. Using the same pretrained models, our warp-guided method is better than the condition method in Vista, which validates that our model can make higher-quality controllable predictions. Fig. 6 presents the visualization comparison of trajectory-based prediction. With the same image input and various trajectories, Vista's results have more distortions and non-compliance with trajectory, while our model can infer high-quality results.

Human Evaluation For the specific driving scenario, given the input trajectory, metrics like FID and FVD can not measure the compliance, so here we introduce human evaluation to judge the ability of trajectory compliance. As shown in Tab. 3b, our method has twice the winning rate of Vista.

Ablation Study

Quantitative Results Tab. 6 shows the effect of the three main modules used in our method, including DiT, coarse-to-fine, and distillation. Decouple learning using coarse-to-fine training brings good improvements in the driving scenario. Structure preservation and distillation continuously improve the video quality. In Tab. 5, compared with two interpolation models (Chen et al. 2023; Wang et al. 2024), our method achieves the best FID and FVD scores, validating the effectiveness of our continuous motion prediction.

Figure 7: Ablation study. The first row shows CDiT predictions including lots of distortions, the second row adds the structure preservation term, and the last row shows the best.

Interpolation Model	FID↓	FVD↓
SEINE (Chen et al. 2023)	23.4	213.6
FRAMER (Wang et al. 2024)	29.0	147.6
Fine-DiT (ours)	**12.3**	**102.9**

Table 5: Frame interpolation evaluation.

DiT	Coarse-to-Fine	$\mathcal{L}_{struct}$	Distill.	FVD↓
✓				347.7
✓	✓			321.8
✓	✓	✓		300.3
✓	✓	✓	✓	**290.6**

Table 6: Ablation study of components for short video prediction evaluated on the validation subset.

Qualitative Results We introduce a structure preservation loss to maintain the quality of each frame. As shown in Fig. 7, comparing the first row and second row, adding the structure preservation term improves the quality and fidelity of each coarse frame, avoiding obvious distortions. We propose the distillation from fine flows to coarse flows to improve the consistency between coarse frames. Comparing the second row and last row in Fig. 7, our distillation improves both the consistency and fidelity of coarse frames.

Generalization Results We additionally compare baselines with our model on side-view images from NuScenes and front-view images from Waymo dataset (Sun et al. 2020). Furthermore, we validate the effectiveness of our distillation framework on the U-Net–based SVD architecture. Additional results are provided in the supplementary materials.

Conclusion

In this paper, we aim to alleviate the error accumulation problem in the long-term driving world model predictions. Intuitive comparisons suggest that the main challenges are 1) the training-inference gap and 2) the gap between general scenes and driving scenes. To address these, we build a simple yet effective long-term driving world model. We decouple long video world model learning into large motion learning and bidirectional continuous motion learning, and leverage scalable DiT to solve them. A novel distillation method is proposed to utilize fine-flow priors to guide the coarse-flow prediction to improve the long video consistency. Extensive experiments on NuScenes demonstrate our model achieving state-of-the-art performance on FVD metrics on various tasks, improving the controllable ability and significantly reducing the inference time for long-term videos.

Acknowledgments

The study was funded by the Shenzhen Science and Technology Program (KQTD20240729102051063), the National Natural Science Foundation of China under contracts No. 62422602, No. 62372010, No. 62425101, No. 62332002, No. 62372010, No. 62206281, and the major key project of the Peng Cheng Laboratory (PCL2021A13 and PCL2025A02). Computing support was provided by Pengcheng Cloudbrain.

References

Bahmani, S.; Skorokhodov, I.; Siarohin, A.; Menapace, W.; Qian, G.; Vasilkovsky, M.; Lee, H.-Y.; Wang, C.; Zou, J.; Tagliasacchi, A.; et al. 2024. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*.

Bao, F.; Xiang, C.; Yue, G.; He, G.; Zhu, H.; Zheng, K.; Zhao, M.; Liu, S.; Wang, Y.; and Zhu, J. 2024. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*.

Bian, W.; Huang, Z.; Shi, X.; Li, Y.; Wang, F.-Y.; and Li, H. 2025. GS-DiT: Advancing Video Generation with Pseudo 4D Gaussian Fields through Efficient Dense 3D Point Tracking. *arXiv preprint arXiv:2501.02690*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 22563–22575.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 11621–11631.

Chen, R.; Wu, Z.; Liu, Y.; Guo, Y.; Ni, J.; Xia, H.; and Xia, S. 2024. UniMLVG: Unified Framework for Multi-view Long Video Generation with Comprehensive Control Capabilities for Autonomous Driving. *arXiv preprint arXiv:2412.04842*.

Chen, X.; Wang, Y.; Zhang, L.; Zhuang, S.; Ma, X.; Yu, J.; Wang, Y.; Lin, D.; Qiao, Y.; and Liu, Z. 2023. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth ICLR*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Ebert, F.; Finn, C.; Dasari, S.; Xie, A.; Lee, A.; and Levine, S. 2018. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*.

Gao, R.; Chen, K.; Xiao, B.; Hong, L.; Li, Z.; and Xu, Q. 2024a. MagicDriveDiT: High-Resolution Long Video Generation for Autonomous Driving with Adaptive Control. *arXiv preprint arXiv:2411.13807*.

Gao, R.; Chen, K.; Xie, E.; Hong, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2023. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*.

Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024b. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. In *NeurIPS (NeurIPS)*.

Guo, J.; Ding, Y.; Chen, X.; Chen, S.; Li, B.; Zou, Y.; Lyu, X.; Tan, F.; Qi, X.; Li, Z.; et al. 2025. DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation. *arXiv preprint arXiv:2503.15208*.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Ha, D.; and Schmidhuber, J. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D.; Lillicrap, T.; Norouzi, M.; and Ba, J. 2020. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*.

Hafner, D.; Pasukonis, J.; Ba, J.; and Lillicrap, T. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2024. CameraCtrl: Enabling Camera Control for Text-to-Video Generation. *arXiv preprint arXiv:2404.02101*.

Henschel, R.; Khachatryan, L.; Hayrapetyan, D.; Poghosyan, H.; Tadevosyan, V.; Wang, Z.; Navasardyan, S.; and Shi, H. 2024. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*.

Hou, C.; Wei, G.; Zeng, Y.; and Chen, Z. 2024. Training-free Camera Control for Video Generation. *arXiv preprint arXiv:2406.10126*.

Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; and Corrado, G. 2023. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*.

Jia, F.; Mao, W.; Liu, Y.; Zhao, Y.; Wen, Y.; Zhang, C.; Zhang, X.; and Wang, T. 2023. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*.

Jiang, J.; Hong, G.; Zhou, L.; Ma, E.; Hu, H.; Zhou, X.; Xiang, J.; Liu, F.; Yu, K.; Sun, H.; et al. 2024. Dive: Dit-based video generation with enhanced control. *arXiv preprint arXiv:2409.01595*.

Kim, S. W.; Philion, J.; Torralba, A.; and Fidler, S. 2021. Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*, 5820–5829.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*.

LeCun, Y. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1).

Li, B.; Guo, J.; Liu, H.; Zou, Y.; Ding, Y.; Chen, X.; Zhu, H.; Tan, F.; Zhang, C.; Wang, T.; et al. 2024. UniScene: Unified Occupancy-centric Driving Scene Generation. *arXiv preprint arXiv:2412.05435*.

Liang, J.; Wu, C.; Hu, X.; Gan, Z.; Wang, J.; Wang, L.; Liu, Z.; Fang, Y.; and Duan, N. 2022. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *NeurIPS*, 35: 15420–15432.

Lu, J.; Huang, Z.; Yang, Z.; Zhang, J.; and Zhang, L. 2025. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *ECCV*, 329–345. Springer.

OpenAI. 2024. Sora technical report. https://openai.com/index/video-generation-models-as-world-simulators/. Accessed: 2024-02-15.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

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

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; Li, Y.; Casas, D. d. L.; Budden, D.; Abdolmaleki, A.; Merel, J.; Lefrancq, A.; et al. 2018. Deepmind control suite. *arXiv preprint arXiv:1801.00690*.

Wang, W.; Wang, Q.; Zheng, K.; Ouyang, H.; Chen, Z.; Gong, B.; Chen, H.; Shen, Y.; and Shen, C. 2024. Framer: Interactive Frame Interpolation. In *The Thirteenth ICLR*.

Wang, X.; and Peng, P. 2025. ProphetDWM: A Driving World Model for Rolling Out Future Actions and Videos. *arXiv preprint arXiv:2505.18650*.

Wang, X.; Wu, C.; Yin, S.; Ni, M.; Wang, J.; Li, L.; Yang, Z.; Yang, F.; Wang, L.; Liu, Z.; Fang, Y.; and Duan, N. 2023a. Learning 3D photography videos via self-supervised diffusion on single images. In *Thirty-Second IJCAI*, IJCAI '23. ISBN 978-1-956792-03-4.

Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; and Lu, J. 2023b. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*.

Wang, Y.; He, J.; Fan, L.; Li, H.; Chen, Y.; and Zhang, Z. 2023c. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *arXiv preprint arXiv:2311.17918*.

Wu, Z.; Ni, J.; Wang, X.; Guo, Y.; Chen, R.; Lu, L.; Dai, J.; and Xiong, Y. 2024. HoloDrive: Holistic 2D-3D Multi-Modal Street Scene Generation for Autonomous Driving. *arXiv preprint arXiv:2412.01407*.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; and Wong, T.-T. 2025. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, 399–417. Springer.

Xu, H.; Peng, P.; Tan, G.; Chang, Y.; Zhao, Y.; and Tian, Y. 2025. Temporal Triplane Transformers as Occupancy World Models. *arXiv preprint arXiv:2503.07338*.

Yang, J.; Gao, S.; Qiu, Y.; Chen, L.; Li, T.; Dai, B.; Chitta, K.; Wu, P.; Zeng, J.; Luo, P.; et al. 2024a. Generalized predictive model for autonomous driving. In *CVPR*, 14662–14672.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Yin, S.; Wu, C.; Yang, H.; Wang, J.; Wang, X.; Ni, M.; Yang, Z.; Li, L.; Liu, S.; Yang, F.; et al. 2023. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*.

Yin, T.; Gharbi, M.; Park, T.; Zhang, R.; Shechtman, E.; Durand, F.; and Freeman, W. T. 2024a. Improved Distribution Matching Distillation for Fast Image Synthesis. *arXiv preprint arXiv:2405.14867*.

Yin, T.; Gharbi, M.; Zhang, R.; Shechtman, E.; Durand, F.; Freeman, W. T.; and Park, T. 2024b. One-step diffusion with distribution matching distillation. In *CVPR*, 6613–6623.

Yin, T.; Zhang, Q.; Zhang, R.; Freeman, W. T.; Durand, F.; Shechtman, E.; and Huang, X. 2024c. From Slow Bidirectional to Fast Causal Video Generators. *arXiv preprint arXiv:2412.07772*.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Zhao, C.; Liu, M.; Wang, W.; Chen, W.; Wang, F.; Chen, H.; Zhang, B.; and Shen, C. 2024. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*.

Zhao, M.; He, G.; Chen, Y.; Zhu, H.; Li, C.; and Zhu, J. 2025. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*.

Zheng, W.; Chen, W.; Huang, Y.; Zhang, B.; Duan, Y.; and Lu, J. 2025. Occworld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*, 55–72. Springer.

Zheng, W.; Xia, Z.; Huang, Y.; Zuo, S.; Zhou, J.; and Lu, J. 2024. Doe-1: Closed-Loop Autonomous Driving with Large World Model. *arXiv preprint arXiv:2412.09627*.