DriveVA: Video Action Models are Zero-Shot Drivers

Mengmeng Liu[1], Diankun Zhang[2], Jiuming Liu[3], Jianfeng Cui[2], Hongwei Xie[2], Guang Chen[2], Hangjun Ye[2], Michael Ying Yang[4], Francesco Nex[1], and Hao Cheng[1]‡

[1]University of Twente, The Netherlands [2]Xiaomi EV, China
[3]University of Cambridge, United Kingdom [4]University of Bath, United Kingdom

Abstract. Generalization is a central challenge in autonomous driving, as real-world deployment requires robust performance under unseen scenarios, sensor domains, and environmental conditions. Recent world-model-based planning methods have shown strong capabilities in scene understanding and multi-modal future prediction, yet their generalization across datasets and sensor configurations remains limited. In addition, their loosely coupled planning paradigm often leads to poor video-trajectory consistency during visual imagination. To overcome these limitations, we propose DriveVA, a novel autonomous driving world model that jointly decodes future visual forecasts and action sequences in a shared latent generative process. DriveVA inherits rich priors on motion dynamics and physical plausibility from well-pretrained large-scale video generation models to capture continuous spatiotemporal evolution and causal interaction patterns. To this end, DriveVA employs a DiT-based decoder to jointly predict future action sequences (trajectories) and videos, enabling tighter alignment between planning and scene evolution. We also introduce a video continuation strategy to strengthen long-duration rollout consistency. DriveVA achieves an impressive PDM-based planning performance of 90.9 PDM score on the NAVSIM benchmark. Extensive experiments also demonstrate the zero-shot capability and cross-domain generalization of DriveVA, which reduces average L2 error and collision rate by 78.9% and 83.3% on nuScenes and 52.5% and 52.4% on the Bench2Drive built on CARLA v2 compared with the state-of-the-art world-model-based planner.

Keywords: World Model · Autonomous Driving · Video Action Models

1 Introduction

Generalization has long been a fundamental goal in autonomous driving, as it is essential for building systems that operate reliably in the real world [9,11,16,18,26,48–51,77]. A capable autonomous driving model should not only perform well on scenarios seen during training, but also remain robust under unseen traffic

‡ Corresponding author.

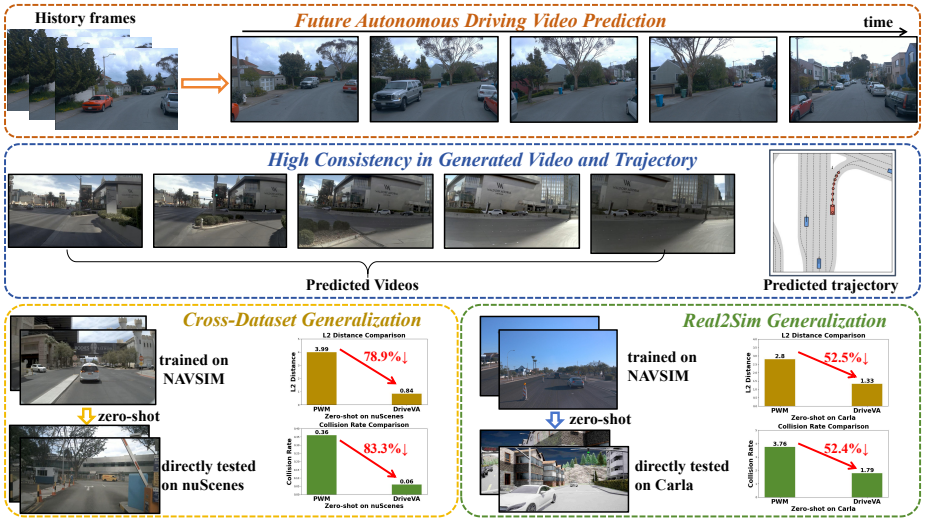

Fig. 1: DriveVA: unified video–trajectory rollout for planning. Given history frames, DriveVA rolls out a future video clip (top). The ego trajectory is generated together with the video rollout and remains aligned with the visual scene evolution (middle). Bottom: zero-shot comparisons trained on NAVSIM and evaluated on nuScenes (cross dataset) and CARLA (cross domain from real to simulation), showing large relative improvements over PWM [83] in displacement error and collision rate.

patterns, novel road layouts, and diverse sensor configurations [28, 36, 90]. This ability is especially important for real-world deployment, where long-tail events, domain shifts, and complex agent interactions are common. Recent advances in large-scale pre-trained models have motivated researchers to develop autonomous driving systems that can better transfer across tasks and domains [42, 43, 74]. This trend has led to the development of Vision-Language-Action (VLA) models [19, 20, 39, 74, 90], which leverage pre-trained vision-language models as the backbone and fine-tune them on driving-specific trajectory data. This strategy can reduce the amount of task-specific training data required while still achieving strong planning performance. However, despite these advances, true generalization, especially zero-shot transfer across datasets, remains limited and has yet to be fully realized. A key reason is that prevailing VLA pretraining on static image–text pairs primarily transfers semantic knowledge ("what is what"), but provides limited spatiotemporal and causal priors ("how the world moves") needed for robust closed-loop planning.

Recently, large-scale video generation models [35, 62, 76, 88] have shown strong generalization to unseen textual prompts and visual contexts. By learning from massive video corpora, they capture realistic motion patterns and physically plausible scene dynamics [6, 35, 62, 68, 69, 81, 82], suggesting rich priors over real-world temporal evolution. Notably, the ability of video generators to produce temporally coherent future predictions under flexible conditioning aligns closely with the goal of building generalizable driving world models. This motivates a

key question: ***"Can large video generation models serve as a foundation for generalizable autonomous driving video action models?"***

To answer this question, we investigate how to build and fine-tune autonomous driving models upon large-scale video generation models. Existing world-model-based planning methods suffer from two major bottlenecks. First, they often exhibit limited generalization across diverse datasets as the world knowledge learned from one dataset is difficult to transfer effectively to another. Second, they commonly suffer from inconsistency between visual and action rollouts, because video imagination and trajectory generation are typically modeled separately or only loosely coupled [67, 80]. To bridge the gap between generic video generation and driving-oriented planning, the key challenge is to enable video generation models not only to synthesize plausible future scenes, but also to produce actionable driving trajectories that can directly guide vehicle planning. Moreover, to effectively transfer the strong generalization capability of video generation models from the visual domain to the planning domain, it is crucial to maintain consistency between the predicted driving trajectories and the visual future evolution represented in the generated videos, as illustrated by the qualitative comparisons in Fig. 3 and Fig. 4. Such alignment allows the semantic understanding and physical priors learned from large-scale video data to be naturally extended to autonomous driving behaviors.

In this paper, we propose a video-action model, called DriveVA, which integrates large-scale video priors with end-to-end planning and dense supervision from world modeling. We find that video-level supervision is the main driver of planning gains, rather than a merely auxiliary loss appended to a pipeline as in most existing methods [39, 67, 78, 87]. Concretely, enabling video supervision boosts NAVSIMv1 PDMS from 71.4 to 90.9 (+19.5) over action-only optimization (Table 5). The key is that video supervision provides dense temporal grounding of scene dynamics, and planning benefits only when the predicted actions are forced to stay consistent with the imagined future. As shown in Fig. 1, this motivates our unified formulation: instead of modeling future visual imagination and trajectory generation in separate stages, DriveVA places future video latents and action tokens in a shared latent generative process and jointly decodes them with a single DiT [56] in a shared latent space, so trajectories are generated as action grounding of the same rollout rather than being optimized in a separate stage. This unified formulation yields tighter video–trajectory alignment. Despite being generative, we observe that as few as two sampling steps already reach near-optimal planning performance, enabling efficient recurrent decision making. We further introduce a video continuation module to maintain long-duration consistency by progressively rolling out future video clips [47, 52]. Extensive experiments demonstrate that DriveVA achieves state-of-the-art PDM-based planning performance on NAVSIM, and also transfers strongly to unseen datasets across real driving scenes (e.g., nuScenes) and simulated scenes (e.g., Bench2Drive) in a zero-shot setting without target-domain fine-tuning, supporting the generalization capability of DriveVA.

Overall, our core contributions are as follows:

- We propose DriveVA, a unified video-action world model for autonomous driving that jointly models future visual imagination and trajectory prediction within a shared latent generative process, alleviating the mismatch caused by cascaded or loosely coupled planning pipelines.
- We design a unified DiT-based decoder that simultaneously generates future video latents and action tokens, leading to stronger video-trajectory consistency and tighter alignment between scene evolution and planned behavior.
- We introduce a video continuation module that progressively rolls out future clips, preserving long-horizon structural consistency during recurrent planning.
- Extensive experiments show that DriveVA achieves state-of-the-art PDM-based planning performance on NAVSIM (**90.9** PDMS), and delivers strong zero-shot performance on nuScenes (trained on NAVSIM) with **78.9%** lower average L2 error and **83.3%** lower collision rate than the state of the art. It also improves generalization from real to simulation on Bench2Drive (CARLA), reducing average L2 error by **52.5%** and collision rate by **52.4%**.

2 Related Work

2.1 Vision Language Action Models

VLAs. Recently, the rapid development of vision-language-action (VLA) models [19, 20, 39, 74, 90] has advanced a new paradigm for language-guided autonomous driving: these models jointly integrate language understanding, environment perception, and vehicle control to accomplish driving tasks in an end-to-end manner. This progress has been largely enabled by the continued evolution of vision [16, 22, 54, 57, 70, 71, 79], language [1, 58], and vision-language [8, 46, 64] foundation models. Despite this progress, most existing driving VLAs are still built upon vision-language models (VLMs) pre-trained on large-scale web data. While such models are effective at transferring visual-semantic knowledge, their pretraining data is primarily composed of *static* image-text pairs, which limits their ability to capture temporal dynamics and physical interaction patterns directly; They do not naturally inherit the spatiotemporal priors required for adapting to new complex interactive scenarios. Consequently, the generalization capability of current driving VLAs, especially when faced with unseen scenarios and unseen behaviors, still exhibits clear limitations [90].

Generalization in VLAs. To address the generalization issue, existing driving VLA methods mainly follow two directions: one focuses on targeted data construction for corner cases, while the other relies on structured expert modeling for long-tail behaviors [10, 26, 28, 90]. However, stronger *zero-shot* generalization remains insufficiently addressed. For example, Impromptu VLA [10] improves robustness through a manually curated corner-case dataset [26, 28], but relies on predefined scenario categories and trajectory-centric supervision, limiting true cross-dataset zero-shot transfer. DriveMoE [74], in contrast, addresses rare and long-tail driving behaviors through scene- and skill-specialized experts [90], yet

still depends on predefined skill partitions and benchmark-specific data distributions, with limited evidence of transfer to unseen platforms or environments. We argue that *zero-shot driving capability* is particularly critical for planning, as it more directly measures whether a model can make reliable decisions when encountering unseen corner cases rather than merely interpolating among observed trajectory patterns, and also serves as an indicator of cross-platform and cross-scenario generalization. In contrast, video-based world models can leverage dense frame-level supervision to learn physical dynamics from visual evolution, offering a more scalable path toward generalization beyond fixed action templates, benchmark-specific skill partitions, and manually defined corner-case taxonomies [21, 39, 65, 73, 87].

2.2 Video Model-based Autonomous Driving

Motivated by intuitive physical reasoning, world models aim to improve driving decisions by forecasting future scene evolution. Existing autonomous-driving world models can be broadly divided into two lines: latent-dynamics models for planning [63, 75, 87], and models that explicitly predict future visual observations for decision making [39, 78, 80].

Latent-dynamics methods, such as LAW [38], World4Drive [87], AdaWM [63], and Raw2Drive [75], learn compact future representations for planning, policy optimization, or robustness, but generally treat the world model as an auxiliary module for supervision or planning guidance rather than using explicit visual rollouts at inference. In contrast, visually predictive approaches, including FutureSightDrive [78], Epona [80], DriveVLA-W0 [39], and DriveLaW [67], more directly exploit future visual prediction for planning. However, existing methods still typically treat visual prediction as an auxiliary signal, an intermediate reasoning process, or a module loosely coupled with planning. Even in methods that connect the two more explicitly, video and action generation are often maintained as separate branches, with consistency relying on inter-module feature transfer or multi-stage optimization. As a result, mismatches between imagined futures and generated actions can accumulate over time, causing the executed actions to deviate from the future evolution predicted by the world model.

Built on video diffusion backbones, recent VAM-style driving models such as VaViM/VaVAM [3] explore video generative modeling for autonomous driving, suggesting that video generation priors can provide useful spatiotemporal and physical priors for driving policies. Unlike latent world models that learn dynamics from scratch in compact latent spaces [2, 23–25], they can directly exploit pretrained video representations that already encode rich physical dynamics. This suggests that a unified generative formulation of future video and actions may provide tighter coupling between visual forecasting and planning, while also improving transfer across data domains. Motivated by this observation, our method adopts a single shared generative process to jointly model future imagination and action generation, and further investigates how data scale and diversity affect generalization in autonomous driving.

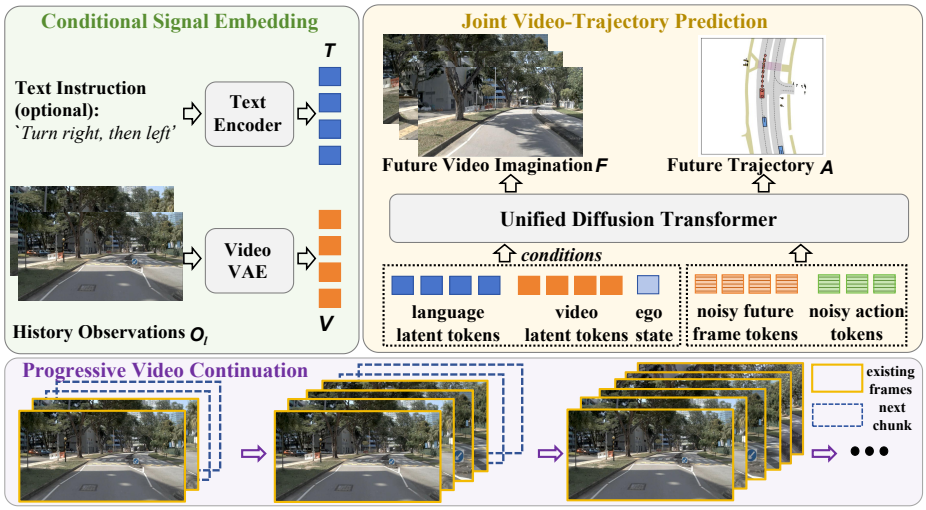

Fig. 2: Overall pipeline of DriveVA. Given history observations, the ego state (current velocity v_x, v_y), and language instructions, the model first encodes conditional signals into latent tokens through a text encoder and a video VAE [62]. A unified diffusion transformer (DiT) [56] then jointly predicts future video latents and future action tokens in a shared generative process, ensuring strong video–trajectory consistency. To maintain long-horizon temporal coherence, a progressive video continuation strategy recursively rolls out future video clips while updating predicted trajectories.

3 Preliminary

Flow Matching. Flow matching [45,53,60] models generation as a continuous-time transformation from a simple source distribution to the target data distribution. Let $x_{\text{data}} \in \mathbb{R}^d$ be a data sample and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a noise sample. The model learns a time-dependent velocity field $v_\theta : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ that defines the dynamics of a trajectory $x^{(s)}$ via

$$\frac{dx^{(s)}}{ds} = v_\theta\left(x^{(s)}, s\right), \qquad x^{(0)} = \boldsymbol{\epsilon}, \quad s \in [0,1]. \tag{1}$$

Intuitively, the learned flow transports samples from noise at $s = 0$ to the data manifold at $s = 1$.

For training, flow matching supervises the model on a prescribed interpolation path between $\boldsymbol{\epsilon}$ and x_{data}. We use the standard linear interpolation $x^{(s)} = (1-s)\boldsymbol{\epsilon} + s x_{\text{data}}$, whose derivative is $\dot{x}^{(s)} = x_{\text{data}} - \boldsymbol{\epsilon}$. The network v_θ is trained to regress this target velocity:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{s,\boldsymbol{\epsilon},x_{\text{data}}}\left[\left\| v_\theta\left(x^{(s)}, s\right) - \dot{x}^{(s)} \right\|_2^2\right]. \tag{2}$$

At inference, generation starts from Gaussian noise and integrates the learned velocity field from $s = 0$ to $s = 1$:

$$x^{(1)} = x^{(0)} + \int_0^1 v_\theta\Big(x^{(s)}, s\Big)\, ds, \qquad x^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

Video Generation with Conditional Flow Matching. Recent video generators [4, 62] commonly perform flow matching in the latent space of a pretrained video autoencoder. Let E and D denote the encoder and decoder, respectively. Given a conditioning signal c, we aim to generate a latent video sequence $\mathbf{z} = \{z_1, \ldots, z_{T_v}\}$ and decode it to pixels with D.

Conditional flow matching learns a velocity field $v_\theta(\mathbf{z}^{(s)}, s \mid c)$ that defines the latent dynamics $\frac{d\mathbf{z}^{(s)}}{ds} = v_\theta(\mathbf{z}^{(s)}, s \mid c)$ with $\mathbf{z}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Integrating from $s = 0$ to 1 yields the clean latent $\mathbf{z}^{(1)}$, which is then decoded by D. This latent-space formulation is efficient and well-suited for long-horizon, condition-controlled video synthesis.

4 Method

4.1 Problem Formulation

Given a language instruction $\mathcal{T}$ (including the high-level command) and a history observation buffer $\mathcal{O}_l = \{\mathbf{F}_{l-m+1}, \ldots, \mathbf{F}_l\}$, which contains m-frame history observations from $\mathbf{F}_{l-m+1}$ to $\mathbf{F}_l$. Our goal at the current timestep l is to jointly predict future actions (trajectories) and future visual imaginations. Specifically, conditioned on $\mathcal{T}$, the current ego state $\mathbf{q}_l$ (represented by the ego velocity components v_x and v_y), and the visual history observations $\mathcal{O}_l$, we predict:

1. **An action chunk** $\mathcal{A}_{l+1:l+K} = \{\boldsymbol{a}_{l+i} \in \mathbb{R}^3\}_{i=1}^K$ consisting of K future actions to be executed sequentially, where each action $\boldsymbol{a}_{l+i}$ is a 3-D vector. The first two dimensions encode the ego-vehicle (x, y) position, and the last dimension encodes the yaw angle.
2. **A future video clip** $\mathcal{F}_{l+1:l+N} = \{\boldsymbol{F}_{l+j}\}_{j=1}^N$ consisting of N frames that depict the anticipated future visual evolution by executing $\mathcal{A}_{l+1:l+K}$. In practice, we do not predict raw frames directly; instead, we predict their latent representations, as detailed in Sec. 4.2.

After executing $\mathcal{A}_{l+1:l+K}$, we obtain new observations, update $\mathcal{O}_l$ using a sliding window, and repeat the process until task completion. This rolling-horizon setup reduces difficulty in long-horizon prediction to a progressive sequence of short *video-continuation* problems.

Joint Video–Action Modeling. Formally, DriveVA jointly models future video imaginations and action chunks conditioned on $\mathbf{C}_l := (\mathcal{O}_l, \mathcal{T}, \mathbf{q}_l)$. This formulation can be viewed as unifying video continuation and IDM-style [17, 89] action grounding within a single end-to-end model, where actions are predicted to be consistent with the imagined future. Instead of training two separate models [37, 55, 67] (a video prediction model and an inverse dynamics model) for

the decomposed objective, we optimize a single model end-to-end with this joint objective. This design encourages tighter video–action alignment through deep cross-modal integration (Fig. 5 and Fig. 4). Moreover, since pretrained video models already provide strong video-prediction priors from large web-scale data, DRIVEVA focuses on adapting these priors to driving-domain video continuation and learning action grounding from predicted visual futures. We further hypothesize that this improves generalization power over conventional VLA training from VLMs, because our formulation explicitly learns temporal dynamics from video frames, which are both used as conditional inputs and prediction targets.

4.2 Data Preprocessing

Text Instruction Encoding. We use a frozen text encoder from Wan2.2-TI2V-5B [62] to encode the language instruction $\mathcal{T}$ (including the high-level command) into a fixed-length token sequence $\mathbf{T} \in \mathbb{R}^{L_T \times d}$ (Fig. 2). These encoded text tokens are injected into the backbone through a cross-attention mechanism, instead of being concatenated with the visual/action stream, keeping the spatiotemporal token sequence compact and decoupling text length for higher control flexibility.

Video Causal VAE with Wan2.2-TI2V-5B. We adopt the 3D-causal VAE encoder from Wan2.2-TI2V-5B as the video encoder. Given a video clip $\mathcal{F} = \{\mathbf{F}_j\}_{j=1}^{N}$, the encoder produces a temporally downsampled latent sequence:

$$\mathcal{V} = \{\mathbf{V}_j\}_{j=1}^{n}, \qquad \mathbf{V}_j \in \mathbb{R}^{h \times w \times c}, \tag{4}$$

where n is the latent sequence length after temporal downsampling. In the original WAN's design, causality ensures that the first latent feature $\mathbf{V}_1$ depends only on the first frame observation $\mathbf{F}_1$, so a single observed frame can be encoded as a valid conditional latent at inference time.

To guarantee long-duration consistency, we further extend this single-frame conditioning to a *video-continuation* setting by conditioning on a history observation buffer rather than only the current frame. Specifically, at current timestep l, we encode the observation buffer $\mathcal{O}_l = \{\mathbf{F}_{l-m+1}, \ldots, \mathbf{F}_l\}$ into a sequence of history latents: $\mathcal{V}_l^{\text{his}} = \{\mathbf{V}_{l-m+1}, \ldots, \mathbf{V}_l\}$. Thus, the first m-frame conditioning latents are all derived from historical observations and provide long-range visual priors for continuation. During training, we encode the full clip to obtain both history latents $\mathcal{V}_l^{\text{his}}$ and future latents $\mathcal{V}_l^{\text{fut}}$; during inference, we encode only $\mathcal{O}_l$ and generate future latents conditioned on the encoded history latents $\mathcal{V}_l^{\text{his}}$.

4.3 Consistent Video-Action Generation

At each current timestep l, DRIVEVA jointly predicts future video latents and action tokens conditioned on (i) history latents $\mathcal{V}_l^{\text{his}}$, (ii) the current ego state $\mathbf{q}_l$, and (iii) text/command tokens $\mathbf{T}$, as in Fig. 2. This design matches training and inference: both use a fixed-length history buffer and predict a short continuation window, which can be chained to progressively generate long-horizon rollout.

Input Tokenization. We raster-flatten each visual latent $\mathbf{V}_t$ in $\mathcal{V}^{\text{his}}$ and project it into the model dimension:

$$\mathbf{V}'_t = \text{Proj}(\text{Flatten}(\mathbf{V}_t)) \in \mathbb{R}^{L_V \times d}, \qquad L_V = h \cdot w. \tag{5}$$

The current ego state $\mathbf{q}_l$ is embedded into L_S state tokens $\mathbf{S}_l \in \mathbb{R}^{L_S \times d}$ using an MLP. Each action $\mathbf{a}_{l+i} \in \mathbb{R}^3$ is also embedded into one token via an MLP, yielding the action-token sequence $\mathbf{A}_{l+1:l+K} \in \mathbb{R}^{K \times d}$.

Fixed Condition and Generative Targets. We split the model input into a *noise-free condition block* and a *generative target block*:

$$\mathbf{X}^{(l)}_{\text{cond}} = [\,\mathbf{S}_l,\ \mathbf{V}'_{l-m+1}, \ldots, \mathbf{V}'_l\,] \in \mathbb{R}^{L_{\text{cond}} \times d}, \tag{6}$$

$$\mathbf{Y}^{(l)}_0 = [\,\mathbf{V}'_{l+1}, \ldots, \mathbf{V}'_{l+n_{\text{pred}}},\ \mathbf{A}_{l+1:l+K}\,] \in \mathbb{R}^{L_{\text{tgt}} \times d}. \tag{7}$$

Here, n_{pred} is the number of future latent steps corresponding to the predicted future clip (after temporal downsampling). The condition block $\mathbf{X}^{(l)}_{\text{cond}}$ is kept fixed at both training and inference. Given the conditional tokens $\mathbf{X}^{(l)}_{\text{cond}}$ and text tokens $\mathbf{T}$, a Diffusion Transformer (DiT) decoder predicts the conditional velocity field for the generative targets:

$$\hat{\mathbf{v}}^{(l,s)}_\theta = f_\theta\Big([\mathbf{X}^{(l)}_{\text{cond}}, \mathbf{Y}^{(l,s)}],\ s \mid \mathbf{T}\Big), \tag{8}$$

where $\mathbf{Y}^{(l,s)}$ is the noisy interpolation of the clean targets $\mathbf{Y}^{(l)}_0$ at flow time s, and f_θ is the DiT decoder parameterized by θ.

4.4 Flow Matching Objective

Following the flow matching formulation in Sec. 3, we instantiate a *conditional* flow over the generative target block $\mathbf{Y}^{(l)}_0$, conditioned on the fixed context block $\mathbf{X}^{(l)}_{\text{cond}}$ and text tokens $\mathbf{T}$.

Flow-Matching Generative Loss. At timestep l, we denote the clean target tokens as $\mathbf{Y}^{(l)}_0$. We sample $s \sim \mathcal{U}(0,1)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and construct the linear interpolation $\mathbf{Y}^{(l,s)} = (1-s)\boldsymbol{\epsilon} + s\mathbf{Y}^{(l)}_0$, whose target velocity is $\dot{\mathbf{Y}}^{(l,s)} = \mathbf{Y}^{(l)}_0 - \boldsymbol{\epsilon}$. We optimize the standard flow-matching regression loss:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{l,s,\mathbf{Y}^{(l)}_0,\boldsymbol{\epsilon}}\left[\left\|\hat{\mathbf{v}}^{(l,s)}_\theta - \dot{\mathbf{Y}}^{(l,s)}\right\|_2^2\right]. \tag{9}$$

5 Experiments

5.1 Datasets

NAVSIM v1. We use the NAVSIM v1 benchmark [14] (built on OpenScene [13]) as our main PDM-based evaluation for safety-critical driving. It reports NC,

Table 1: Performance comparison on NAVSIM *Navtest* using PDM-based metrics. Methods are grouped by whether they employ an explicit world model: *Traditional End-to-End Methods* and *World Model Methods*.

Method	Ref	Image	Lidar	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
Constant Velocity	-			68.0	57.8	50.0	**100**	19.4	20.6
Ego Status MLP [14]	arXiv'23			93.0	77.3	83.6	**100**	62.8	65.6
Traditional End-to-End Methods									
VADv2-$\mathcal{V}_{8192}$ [33]	ICLR'26	✓		97.2	89.1	91.6	**100**	76.0	80.9
UniAD [31]	CVPR'23	✓		97.8	91.9	92.9	**100**	78.8	83.4
TransFuser [12]	TPAMI'23	✓	✓	97.7	92.8	92.8	**100**	79.2	84.0
PARA-Drive [66]	CVPR'24	✓		97.9	92.4	93.0	99.8	79.3	84.0
ReCogDrive-IL [41]	ICLR'26	✓		98.1	94.7	94.2	**100**	80.9	86.5
DiffusionDrive [44]	CVPR'25	✓	✓	98.2	96.2	94.7	**100**	82.2	88.1
World Model Methods									
DrivingGPT [7]	ICCV'25	✓		98.9	90.7	94.9	95.6	79.7	82.4
LAW [38]	ICLR'25	✓		96.4	95.4	88.7	99.9	81.7	84.6
Epona [80]	ICCV'25	✓		97.9	95.1	93.8	99.9	80.4	86.2
Resim [72]	NeurIPS'25	✓		–	–	–	–	–	86.6
WoTE [40]	ICCV'25	✓	✓	98.5	96.8	94.9	99.9	81.9	88.3
DriveVLA-W0 [39]	ICLR'26	✓		98.4	95.3	95.2	**100**	80.9	87.2
PWM [83]	NeurIPS'25	✓		98.6	95.9	95.4	**100**	81.8	88.1
Ours	-	✓		**99.2**	**97.5**	**98.7**	**100**	**83.5**	**90.9**

DAC, TTC, Comfort (C.), and Ego Progress (EP), aggregated as PDMS = NC × DAC × $\frac{5\text{EP}+5\text{TTC}+2\text{C.}}{12}$.

nuScenes. For cross-dataset zero-shot evaluation, we evaluate on the nuScenes validation split (150 scenes) from the 1,000-scene nuScenes dataset [5], and report Displacement Error (DE) and Collision Rate (CR) following prior works [31,34].

Bench2Drive. Bench2Drive [32] is a CARLA v2 closed-loop benchmark [15] with diverse interactive scenarios and evaluation routes. We evaluate: (1) From real to simulation cross domain zero-shot transfer by testing a NAVSIM-trained model directly on the Bench2Drive validation split; (2) *sim-enhanced* training by mixing NAVSIM and Bench2Drive data, then evaluating on NAVSIM. Note that transferring policies across real-world logs and simulation is challenging due to the well-known *reality gap* in appearance, dynamics, and agent behaviors [30].

5.2 Training Details

We utilize Wan2.2-TI2V-5B [62] as our pre-trained backbone. Each training sample consists of 4 history frames and 8 future frames at 2 FPS with a resolution of 832 × 480. Training is performed on NVIDIA H20 GPUs with AdamW using a learning rate of 10^{-4} and weight decay of 0.01 under distributed bf16 mixed-precision training. We first train with a batch size of 80 for 20k steps for faster convergence, and then continue fine-tuning for another 10k steps with an effective batch size of 640 via gradient accumulation. We adopt a linear warm-up over the first $1k$ steps starting from 10^{-3} of the base learning rate, followed by a constant learning rate schedule. The training objective combines a flow-matching loss for future-frame generation with a trajectory prediction loss. During inference, we use 2 sampling steps for flow-based sampling.

Table 2: Zero-shot end-to-end motion planning performance on nuScenes [5] and Bench2Drive (CARLA) [32]. All methods are trained on NAVSIM and directly evaluated on the target datasets without target-domain fine-tuning.

Method	Finetune	Ref	nuScenes								Bench2Drive (CARLA)							
			L2 (m) ↓				Collision (%) ↓				L2 (m) ↓				Collision (%) ↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
VLA-World Model Methods																		
DriveVLA-W0 [39]	✗	ICLR'26	0.43	1.26	2.60	1.43	0.22	0.66	1.42	0.77	1.01	2.77	5.22	3.00	1.49	2.53	3.53	2.52
World Model Methods																		
PWM [83]	✗	NeurIPS'25	2.06	3.91	6.00	3.99	0.12	0.15	0.86	0.36	1.70	2.74	3.97	2.80	4.01	3.73	3.53	3.76
Ours	✗	-	**0.33**	**0.76**	**1.43**	**0.84**	**0.00**	**0.07**	**0.12**	**0.06**	**0.69**	**1.29**	**2.03**	**1.33**	**1.38**	**1.97**	**2.65**	**1.79**

Table 3: End-to-end motion planning performance on nuScenes [5] dataset. * represents only using the front camera as input.

Method	nuScenes Finetune	Ref	Input	Auxiliary Supervision	L2 (m) ↓				Collision Rate (%) ↓			
					1s	2s	3s	Avg.	1s	2s	3s	Avg.
ST-P3 [29]	✓	ECCV'22	Camera	Map&Box&Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [31]	✓	CVPR'23	Camera	Map&Box&Motion	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
OccNet [61]	✓	ICCV'23	Camera	3D-Occ&Map&Box	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
OccWorld [84]	✓	ECCV'24	Camera	3D-Occ	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87
VAD-Tiny [34]	✓	ICCV'23	Camera	Map&Box&Motion	0.60	1.23	2.06	1.30	0.31	0.53	1.33	0.72
VAD-Base [34]	✓	ICCV'23	Camera	Map&Box&Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
GenAD [85]	✓	ECCV'24	Camera	Map&Box&Motion	0.36	0.83	1.55	0.91	0.06	0.23	1.00	0.43
Doe-1 [86]	✓	arXiv'24	Camera*	QA	0.50	1.18	2.11	1.26	0.04	0.37	1.19	0.53
Epona [80]	✓	ICCV'25	Camera*	None	0.61	1.17	1.98	1.25	0.01	0.22	0.85	0.36
Ours	✗	-	Camera*	None	**0.33**	**0.76**	**1.43**	**0.84**	**0.00**	**0.07**	**0.12**	**0.06**

5.3 Quantitative Comparison Results

Comparison Results on NAVSIM v1. As shown in Table 1, we evaluate our DriveVA on the NAVSIM dataset using PDM-based metrics. As for the average PDMS metric, our method not only surpasses previous traditional end-to-end methods like DiffusionDrive [44], but also outperforms recent world model-based methods, including latent-dynamics approaches such as LAW [38] and visually predictive models that explicitly forecast future observations for decision making, e.g., Epona [80], PWM [83] and DriveVLA-W0 [39], etc. Notably, compared to some recent methods such as WoTE [40] using multi-modal information as inputs, our method uses only front-view camera images, but still achieves a better score on the PDM-based planning benchmark. We attribute these gains primarily to our unified video-action formulation, which jointly models future video imagination and ego trajectory prediction within a shared generative process, leading to better alignment between what the model imagines and how it plans.

5.4 Cross-Domain Generalization

Zero-shot Evaluation. Table 2 evaluates strict zero-shot transfer, where all methods are trained on NAVSIM and directly evaluated on nuScenes and Bench2Drive without fine-tuning. On nuScenes, DriveVA achieves the best performance among world-model methods across all horizons, with an average L2 error of 0.84 and an average collision rate of 0.06. Under the same protocol, camera-only input, and without auxiliary supervision, it substantially improves

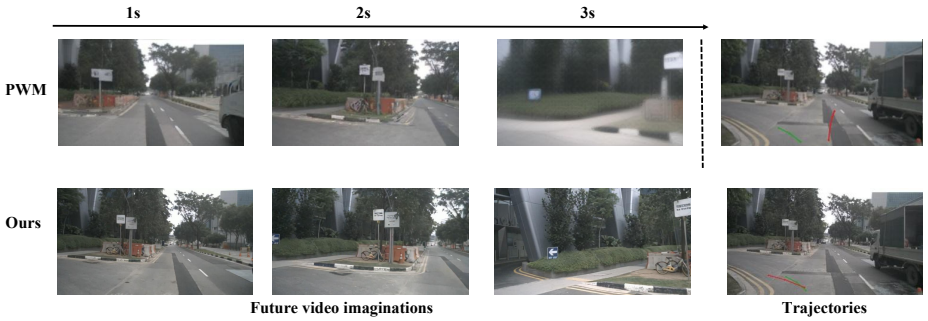

Fig. 3: Video–trajectory consistency comparison in zero-shot unseen nuScenes scenarios. In this left-turn scenario, our method produces trajectories that follow the scene evolution in the generated future video. In contrast, PWM [83] predicts a straight trajectory while the generated video indicates a left-turn maneuver, revealing a clear video–trajectory mismatch. Here, green denotes the GT trajectory and red the predicted trajectory.

over PWM [83], reducing the average L2 error from 3.99 to 0.84 (78.9%) and the average collision rate from 0.36 to 0.06 (83.3%).

On Bench2Drive, DriveVA also shows strong robustness under real-to–simulation transfer. Compared with PWM, it lowers the average L2 error from 2.80 to 1.33 and the average collision rate from 3.76 to 1.79. It further outperforms DriveVLA-W0 [39], reducing the average L2 error from 3.00 to 1.33 (55.7%) and the average collision rate from 2.52 to 1.79 (29.0%). Moreover, Table 3 shows that DriveVA, even without nuScenes training or fine-tuning, surpasses baselines trained or fine-tuned on nuScenes, indicating that unified video-action modeling learns transferable planning priors rather than relying on target-domain adaptation.

5.5 Video–Trajectory Consistency Analysis

To verify video–trajectory consistency, we reconstruct ego motion from ground-truth and generated future videos using DPVO [59]. After aligning each reconstructed trajectory to its reference with a 2D similarity transform, we compute the average L2 error over the future 4s horizon. Table 4 shows consistently low errors on NAVSIM and zero-shot nuScenes, indicating that generated videos imply motion aligned with the predicted trajectories. Fig. 4 further shows that DPVO reconstructions from generated videos closely follow model predictions across representative scenarios. More details and analyses are provided in Sec. S4 of the supplementary material.

Why does DriveVA generalize better in zero-shot setting? Our zero-shot visualizations in Fig. 3 provide a clear explanation: DriveVA maintains strong *video–trajectory consistency* under domain shift, where the predicted trajectory aligns well with the imagined future scene evolution. In contrast, PWM often exhibits noticeable video–trajectory mismatch in zero-shot transfer, leading

Table 4: Quantitative video–trajectory consistency measured by DPVO reconstruction. We run DPVO [59] on ground-truth future videos and generated future videos to reconstruct camera trajectories. After 2D similarity alignment, we compute the average L2 error over the future 4s horizon with respect to the corresponding reference trajectories. Lower is better.

Split / Scenario	GT traj. vs. GT-video recon. Avg. L2 (4s) ↓	Pred. traj. vs. Pred.-video recon. Avg. L2 (4s) ↓
NAVSIM	0.09	0.16
nuScenes	0.07	0.14
Average	0.08	0.15

Table 5: Ablation studies. "Video Loss" indicates video supervision during training. "CARLA Mix Training" denotes joint training with CARLA simulated data and NAVSIM data. "Video Continuation" refers to video-to-video generation.

ID	Video Loss	CARLA Mix Training	Video Continuation	Planning Metric					
				NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
1	✗	✓	✓	95.0	89.0	93.9	86.6	59.7	71.4
2	✓	✗	✓	99.0	97.3	98.4	100	83.2	90.5
3	✓	✓	✗	94.9	95.6	94.2	100	76.9	84.6
4	✓	✓	✓	**99.2**	**97.5**	**98.7**	**100**	**83.5**	**90.9**

Table 6: Future video frames.

Future Frames	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
4	96.6	91.4	95.5	93.3	77.2	82.1
8	**99.2**	**97.5**	**98.7**	**100**	**83.5**	**90.9**
12	98.6	94.4	97.5	99.8	79.5	86.7

Table 7: Training strategy.

Training Strategy	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
From Scratch	89.9	76.8	87.6	99.9	76.8	62.9
LoRA Fine-tune	92.4	88.0	91.0	99.9	67.5	74.9
Full Fine-tune	**99.2**	**97.5**	**98.7**	**100**	**83.5**	**90.9**

Table 8: Sampling steps.

Steps	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
1	61.8	36.9	50.3	1.9	36.9	13.2
2	**99.2**	**97.5**	**98.7**	**100**	83.5	**90.9**
3	99.1	97.4	**98.7**	**100**	**83.7**	**90.9**

Table 9: Model size.

Model Size	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
5B LoRA	92.4	88.0	91.0	99.9	67.5	74.9
14B LoRA	96.3	91.3	95.7	99.4	71.6	80.6
5B Full Fine-tune	**99.2**	**97.5**	**98.7**	**100**	**83.5**	**90.9**

the planned motion to deviate from the future states implied by its imagination and resulting in error accumulation in recurrent rollout. For instance, in Fig. 3, PWM imagines a left-turning future, yet predicts a near-straight trajectory, revealing a clear inconsistency between its visual rollout and planned motion.

Simulation-Enhanced Real-World Transferring. DriveVA further benefits from joint training with simulation data, as shown in Table 5 (ID 2 vs. 4). Simulated environments such as CARLA provide diverse, high-quality corner-case scenarios that are hard to obtain at scale in real-world datasets, making them a valuable source of transferable driving priors. By jointly training on NAVSIM and simulation data, DriveVA achieves improved planning performance in TTC (98.7) and PDMS (90.9) and maintains high performance in other planning metrics on real-world benchmarks, showing that simulation can enhance real-world planning.

5.6 Qualitative Analysis

The qualitative results in Fig. 5 further support the quantitative findings. Our DriveVA generates future video imaginations that remain visually coherent

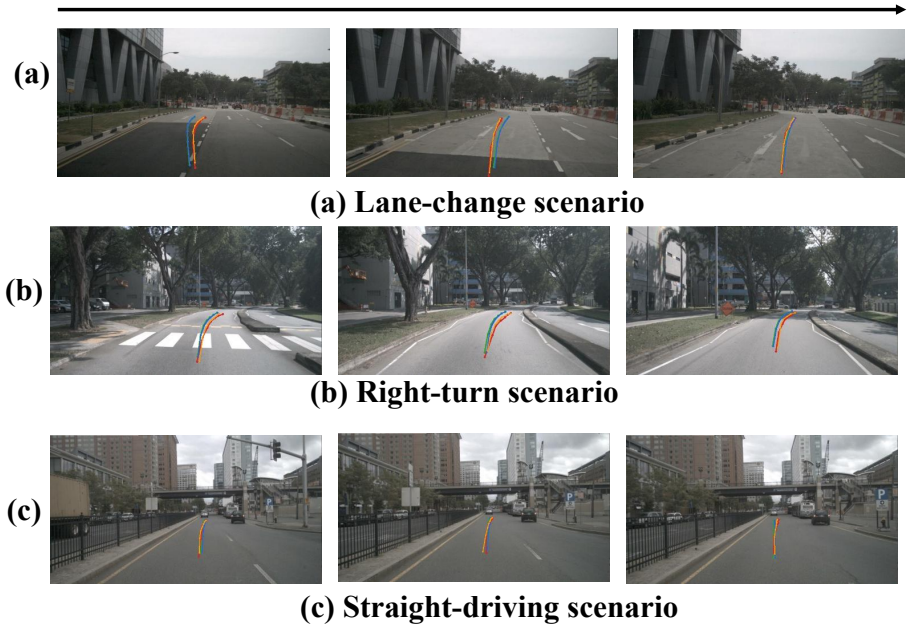

Fig. 4: DPVO-based qualitative analysis of video–trajectory consistency on nuScenes. We visualize zero-shot nuScenes scenarios with temporal frames, including lane-change, right-turn, and straight-driving cases. GT Future and Pred Future denote the ground-truth and predicted trajectories, while DPVO(gt img) and DPVO(pred img) denote DPVO reconstructions from the ground-truth and predicted future videos. The close alignment among these curves provides qualitative evidence of video–trajectory consistency in DriveVA.

over time while producing trajectories that stay well aligned with the evolving scene content. This strong video-trajectory consistency is a direct consequence of our unified generation design: instead of predicting visual futures and actions in separate stages, DriveVA jointly decodes them within a shared latent generative process. As a result, the predicted trajectory more faithfully follows the semantic layout and motion trends implied by the generated future frames. DriveVA not only achieves better planning metrics, but also produces more trustworthy future imaginations whose spatial evolution remains aligned with the intended driving behavior.

5.7 Ablation Study

Effect of key designs. Table 5 shows that video supervision is the main driver of the gain: removing Video Loss drops PDMS from 90.9 to 71.4. Removing Video Continuation also degrades PDMS to 84.6, indicating that the video–action coupling should be preserved during rollout. CARLA mix training further improves PDMS from 90.5 to 90.9, showing the benefit of simulation data.

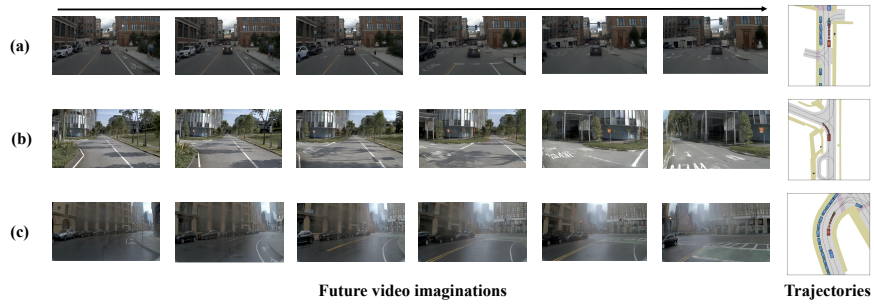

Fig. 5: Visualization of predicted video imaginations and corresponding trajectories. The predicted trajectories follow the scene evolution in the generated future video frames, demonstrating strong video–trajectory consistency enabled by our unified generation framework.

Sampling steps. Table 8 shows that one sampling step is insufficient, while two steps already reach 90.9 PDMS and three steps bring no additional gain. This indicates that DriveVA can perform efficient planning with very few flow-matching steps.

Prediction time horizon. Table 6 varies the video rollout length while fixing the trajectory horizon to K=8 (4s). Eight future frames perform best, whereas shorter rollouts insufficiently cover the action chunk and longer rollouts introduce additional drift.

Training strategy. Table 7 compares training from scratch, LoRA [27], and full fine-tuning. Full fine-tuning performs best, supporting that effective transfer requires adapting the video prior under joint video-level supervision.

Model size. Table 9 further compares different model scales and tuning strategies, showing that full fine-tuning the 5B backbone performs best.

6 Conclusion

In this paper, we propose DriveVA, a unified video–action world model for autonomous driving. DriveVA jointly generates future video latents and trajectory tokens within a shared conditional generative process, improving video–trajectory consistency in real world rollout. Built on a large pretrained video generation backbone and further enhanced with progressive video continuation, DriveVA achieves state-of-the-art PDM-based performance on NAVSIM and shows strong zero-shot transfer to nuScenes and Bench2Drive without target-domain fine-tuning, which provides a new insight in this research domain.

Acknowledgements

This work has been supported by the Centre for Spatial Intelligence (RCSI) at University of Bath, the European Union under grant agreement no. 101136006-XTREME, and the European Innovation Council under grant agreement no. 101257536-CEREBRIS.

References

1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., et al.: V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Bartoccioni, F., Ramzi, E., Besnier, V., Venkataramanan, S., Vu, T.H., Xu, Y., Chambon, L., Gidaris, S., Odabas, S., Hurych, D., et al.: Vavim and vavam: Autonomous driving through video generative modeling. arXiv preprint arXiv:2502.15672 (2025)
4. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators. OpenAI Blog (2024), https://openai.com/research/video-generation-models-as-world-simulators, accessed: 30 Jun. 2026
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
6. Chen, H., Zhou, K., Hua, H., Zhang, K., Qian, J., Ma, W., Chen, H., Liu, C., Zhao, Y., Wang, X., Li, W., Yuille, A., Liang, P.P., Du, Y.: Memobench: Benchmarking world modeling in dynamically changing environments. arXiv preprint arXiv:2606.27537 (2026)
7. Chen, Y., Wang, Y., Zhang, Z.: Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26890–26900 (2025)
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
9. Cheng, H., Liu, M., Chen, L., Broszio, H., Sester, M., Yang, M.Y.: Gatraj: A graph- and attention-based multi-agent trajectory prediction model. ISPRS Journal of Photogrammetry and Remote Sensing **205**, 163–175 (2023)
10. Chi, H., Gao, H.a., Liu, Z., Liu, J., Liu, C., Li, J., Yang, K., Yu, Y., Wang, Z., Li, W., et al.: Impromptu vla: Open weights and open data for driving vision-language-action models. arXiv preprint arXiv:2505.23757 (2025)
11. Chi, H., Qiu, D., Su, H., Liu, H., Li, Z., Zhang, H., Lv, C.: Driver-wm: A driver-centric traffic-conditioned latent world model for in-cabin dynamics rollout. arXiv preprint arXiv:2605.05092 (2026)
12. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A.: Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE transactions on pattern analysis and machine intelligence **45**(11), 12878–12895 (2022)
13. Contributors, O.: Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, Vancouver, Canada. pp. 18–22 (2023)
14. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., et al.: Navsim: Data-driven non-reactive

autonomous vehicle simulation and benchmarking. Advances in Neural Information Processing Systems **37**, 28706–28719 (2024)
15. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
16. Du, S., Zou, Y., Wang, Z., Li, X., Li, Y., Shang, C., Shen, Q.: Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. International Journal of Computer Vision **134**(4), 152 (2026)
17. Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., Abbeel, P.: Learning universal policies via text-guided video generation. Advances in neural information processing systems **36**, 9156–9172 (2023)
18. Feng, T., Wang, W., Yang, Y.: A survey of world models for autonomous driving. arXiv preprint arXiv:2501.11260 (2025)
19. Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., Bai, X.: Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24823–24834 (2025)
20. Fu, H., Zhang, D., Zhao, Z., Cui, J., Xie, H., Wang, B., Chen, G., Liang, D., Bai, X.: Minddrive: A vision-language-action model for autonomous driving via online reinforcement learning. arXiv preprint arXiv:2512.13636 (2025)
21. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems **37**, 91560–91596 (2024)
22. Guo, Z., Duan, C., Guan, T., Wang, Z., Zhou, K., Yan, P.: Vitexqa: A multi-frame temporal perception dataset for video text question answering. arXiv preprint arXiv:2606.24602 (2026)
23. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)
24. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. arXiv preprint arXiv:2010.02193 (2020)
25. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2023)
26. Hao, Y., Li, Z., Sun, L., Wang, W., Yi, N., Song, S., Qin, C., Zhou, M., Zhan, Y., Lang, X.: Driveaction: A benchmark for exploring human-like driving decisions in vla models. arXiv preprint arXiv:2506.05667 (2025)
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)
28. Hu, H., Zuo, J., Lou, Y., Cui, Y., Wang, J., Guan, N., Wang, J., Li, Y.H., Xue, C.J.: Vlm-c4l: Continual core dataset learning with corner case optimization via vision-language models for autonomous driving. arXiv preprint arXiv:2503.23046 (2025)
29. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: European Conference on Computer Vision. pp. 533–549. Springer (2022)
30. Hu, X., Li, S., Huang, T., Tang, B., Huai, R., Chen, L.: How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence. IEEE Transactions on Intelligent Vehicles **9**(1), 593–612 (2023)
31. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)

32. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. Advances in Neural Information Processing Systems **37**, 819–844 (2024)
33. Jiang, B., Chen, S., Gao, H., Liao, B., Zhang, Q., Liu, W., Wang, X.: VADv2: End-to-end vectorized autonomous driving via probabilistic planning. In: The Fourteenth International Conference on Learning Representations (2026)
34. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
35. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
36. Li, J., Xu, R., Ma, J., Zou, Q., Ma, J., Yu, H.: Domain adaptive object detection for autonomous driving under foggy weather. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 612–622 (2023)
37. Li, L., Zhang, Q., Luo, Y., Yang, S., Wang, R., Han, F., Yu, M., Gao, Z., Xue, N., Zhu, X., Shen, Y., Xu, Y.: Causal world modeling for robot control. arXiv preprint arXiv:2601.21998 (2026)
38. Li, Y., Fan, L., He, J., Wang, Y., Chen, Y., Zhang, Z., Tan, T.: Enhancing end-to-end autonomous driving with latent world model. arXiv preprint arXiv:2406.08481 (2024)
39. Li, Y., Shang, S., Liu, W., Zhan, B., Wang, H., Wang, Y., Chen, Y., Wang, X., An, Y., Tang, C., et al.: Drivevla-w0: World models amplify data scaling law in autonomous driving. arXiv preprint arXiv:2510.12796 (2025)
40. Li, Y., Wang, Y., Liu, Y., He, J., Fan, L., Zhang, Z.: End-to-end driving with online trajectory evaluation via bev world model. arXiv preprint arXiv:2504.01941 (2025)
41. Li, Y., Xiong, K., Guo, X., Li, F., Yan, S., Xu, G., Zhou, L., Chen, L., Sun, H., Wang, B., et al.: Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. arXiv preprint arXiv:2506.08052 (2025)
42. Liang, G., Hu, J., Xing, X., Zhang, J., Yu, Q.: Multi-object sketch animation with grouping and motion trajectory priors. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9237–9246 (2025)
43. Liang, G., Wang, Z., Hu, J., Zhou, H., Xue, Z., Zhang, J., Xu, D., Yu, Q.: Render-in-the-loop: Vector graphics generation via visual self-feedback. arXiv preprint arXiv:2604.20730 (2026)
44. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., et al.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12037–12047 (2025)
45. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
46. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
47. Liu, J., Liu, M., Zhu, S., Zhang, Y., Li, J., Yang, M.Y., Nex, F., Cheng, H., Wang, H.: Arflow: Auto-regressive optical flow estimation for arbitrary-length videos via progressive next-frame forecasting. In: The Fourteenth International Conference on Learning Representations (2026)

48. Liu, J., Wang, G., Liu, Z., Jiang, C., Pollefeys, M., Wang, H.: Regformer: An efficient projection-aware transformer network for large-scale point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8451–8460 (2023)
49. Liu, J., Wang, G., Ye, W., Jiang, C., Han, J., Liu, Z., Zhang, G., Du, D., Wang, H.: Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15109–15119 (2024)
50. Liu, M., Cheng, H., Chen, L., Broszio, H., Li, J., Zhao, R., Sester, M., Yang, M.Y.: Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2039–2049 (2024)
51. Liu, M., Liu, J., Yang, M.Y., Jiang, C., Li, J., Zhang, Y., Wang, H., Nex, F., Cheng, H.: Streamvlo: Streaming visual-lidar odometry with cumulative drift compensation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 39086–39097 (2026)
52. Liu, M., Liu, J., Zhang, Y., Li, J., Yang, M.Y., Nex, F., Cheng, H.: 4dstr: Advancing generative 4d gaussians with spatial-temporal rectification for high-quality and consistent 4d generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 7224–7232 (2026)
53. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
54. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
55. Pai, J., Achenbach, L., Montesinos, V., Forrai, B., Mees, O., Nava, E.: mimic-video: Video-action models for generalizable robot control beyond vlas. arXiv preprint arXiv:2512.15692 (2025)
56. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
58. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
59. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. Advances in Neural Information Processing Systems **36**, 39033–39051 (2023)
60. Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., Bengio, Y.: Improving and generalizing flow-based generative models with minibatch optimal transport. Transactions on Machine Learning Research pp. 1–34 (2024)
61. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8406–8415 (2023)
62. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

63. Wang, H., Ye, X., Tao, F., Pan, C., Mallik, A., Yaman, B., Ren, L., Zhang, J.: Adawm: Adaptive world model based planning for autonomous driving. arXiv preprint arXiv:2501.13072 (2025)
64. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
65. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14749–14759 (2024)
66. Weng, X., Ivanovic, B., Wang, Y., Wang, Y., Pavone, M.: Para-drive: Parallelized architecture for real-time autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15449–15458 (2024)
67. Xia, T., Li, Y., Zhou, L., Yao, J., Xiong, K., Sun, H., Wang, B., Ma, K., Chen, G., Ye, H., et al.: Drivelaw: Unifying planning and video generation in a latent driving world. arXiv preprint arXiv:2512.23421 (2025)
68. Xiang, X., Chen, Y., Zhang, G., Wang, Z., Gao, Z., Xiang, Q., Shang, G., Liu, J., Huang, H., Gao, Y., et al.: Macro-from-micro planning for high-quality and parallelized autoregressive long video generation. arXiv preprint arXiv:2508.03334 (2025)
69. Xiang, X., Duan, Z., Zhang, G., Zhang, H., Gao, Z., Wu, J., Zhang, S., Wang, T., Fan, Q., Guo, C.: Pathwise test-time correction for autoregressive long video generation. arXiv preprint arXiv:2602.05871 (2026)
70. Xiao, X., Zhang, Y., Li, X., Wang, T., Wang, X., Wei, Y., Hamm, J., Xu, M.: Visual instance-aware prompt tuning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 2880–2889 (2025)
71. Xiao, X., Zhang, Y., Zhao, L., Liu, Y., Liao, X., Mai, Z., Li, X., Wang, X., Xu, H., Hamm, J., Lin, X., Xu, M., Wang, Q., Wang, T., Han, C.: Prompt-based adaptation in large-scale vision models: A survey. Transactions on Machine Learning Research (2026)
72. Yang, J., Chitta, K., Gao, S., Chen, L., Shao, Y., Jia, X., Li, H., Geiger, A., Yue, X., Chen, L.: Resim: Reliable world simulation for autonomous driving. arXiv preprint arXiv:2506.09981 (2025)
73. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., et al.: Generalized predictive model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14662–14672 (2024)
74. Yang, Z., Chai, Y., Jia, X., Li, Q., Shao, Y., Zhu, X., Su, H., Yan, J.: Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. arXiv preprint arXiv:2505.16278 (2025)
75. Yang, Z., Jia, X., Li, Q., Yang, X., Yao, M., Yan, J.: Raw2drive: Reinforcement learning with aligned world models for end-to-end autonomous driving (in carla v2). arXiv preprint arXiv:2505.16394 (2025)
76. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
77. Yu, H., Jin, Y., Canevaro, A., Schmidt, J., Jordan, J., Li, P., Kaufeld, M., Lindner, S., Betz, J., Stork, W.: G2dp: Diffusion planning with spatio-temporal grid guidance. arXiv preprint arXiv:2606.26017 (2026)

78. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X., Guo, N.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv:2505.17685 (2025)
79. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
80. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., et al.: Epona: Autoregressive diffusion world model for autonomous driving. arXiv preprint arXiv:2506.24113 (2025)
81. Zhao, C., Chen, J., Li, H., Kang, Z., Lu, S., Wei, X., Zhang, K., Yang, J., Tai, Y.: LUVE : Latent-cascaded ultra-high-resolution video generation with dual frequency experts. In: Forty-third International Conference on Machine Learning (2026)
82. Zhao, Y., Wang, Y., Wang, X., Wu, Y., Zhang, H., Haji-Ali, M., Abdal, R., Mirzaei, A., Li, Y., Menapace, W., et al.: Geostream: Toward precise camera controlled streaming video generation. arXiv preprint arXiv:2606.15162 (2026)
83. Zhao, Z., Fu, T., Wang, Y., Wang, L., Lu, H.: From forecasting to planning: Policy world model for collaborative state-action prediction. arXiv preprint arXiv:2510.19654 (2025)
84. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
85. Zheng, W., Song, R., Guo, X., Zhang, C., Chen, L.: Genad: Generative end-to-end autonomous driving. In: European Conference on Computer Vision. pp. 87–104. Springer (2024)
86. Zheng, W., Xia, Z., Huang, Y., Zuo, S., Zhou, J., Lu, J.: Doe-1: Closed-loop autonomous driving with large world model. arXiv preprint arXiv:2412.09627 (2024)
87. Zheng, Y., Yang, P., Xing, Z., Zhang, Q., Zheng, Y., Gao, Y., Li, P., Zhang, T., Xia, Z., Jia, P., et al.: World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28632–28642 (2025)
88. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
89. Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.Y., Gan, C.: Robodreamer: Learning compositional world models for robot imagination. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 235 (2024)
90. Zhou, X., Han, X., Yang, F., Ma, Y., Tresp, V., Knoll, A.: Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13782–13790 (2026)