INWORLD: AN INSTANT-INTERACTIVE MULTIMODAL WORLD MODEL FOR AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

World models have become a scalable alternative to real-world interaction for policy learning and evaluation in autonomous driving. A high-performance driving world model must satisfy three core requirements simultaneously: cross-modal consistency, real-time generation, and pose-controllable interaction. However, existing methods struggle to fulfill all these criteria, frequently requiring trade-offs among them. We present **InWorld**, an interactive driving world model that synchronously generates cross-modally consistent images and LiDAR point clouds in real time. To faithfully capture intrinsic LiDAR sensor characteristics, we design a dedicated LiDAR VAE optimized for range image representation, which integrates horizontal circular convolution and a differentiable ray-drop module. For cross-modally consistent generation, we present a layout architecture equipped with epipolar masked attention, alongside image-LiDAR consistent Plücker coordinates encoding. To enable controllable interactive modeling, we propose temporal decay KV caching and embed it into the teacher-student pipeline via distribution matching distillation. Extensive experiments on the nuScenes dataset demonstrate that InWorld achieves state-of-the-art performance in both generation efficiency and cross-modal alignment, while the student enables interactive generation at a 10 Hz streaming inference rate. Our code will be open-sourced, with the anonymized version at: `https://anonymous.4open.science/r/InWorld-784A`.

1 INTRODUCTION

World models Agarwal et al. (2025); Zhao et al. (2025a); Team et al. (2025); Ren et al. (2025); Ni et al. (2025); Robbyant Team et al. (2026) serve as a promising alternative to real-world interaction, enabling scalable, controllable simulation environments for autonomous driving. These models can generate controllable virtual scenarios to facilitate end-to-end policy learning and evaluation. As concluded in previous literature Ding et al. (2025); Feng et al. (2025), a practical driving world model must meet three essential criteria: real-time generation, cross-modal consistency, and precise pose-controllable interaction. Nevertheless, existing methods can hardly optimize all criteria simultaneously, leading to unavoidable inherent trade-offs across these key capabilities.

Prior work generally falls into three categories. *Vision-only models* Wang et al. (2024); Zhao et al. (2025b); Gao et al. (2024a; 2025); Zhang et al. (2025) produce high-quality visual videos but lack accurate geometric constraints. *Multimodal models* Zhao et al. (2026); Zhang et al. (2024); Bogdoll et al. (2023) possess generation capability with visual and geometric consistency, but most are designed for offline generation and cannot support real-time interactive control. *Interactive models* Savva et al. (2026); He et al. (2025b); Wang et al. (2026) realize fast action-conditioned synthesis, while they lack driving-specific adaptation and precise pose control for autonomous scenarios. These limitations lead to three critical drawbacks for closed-loop autonomous driving: (1) *incomplete cross-modal consistency*, (2) *low generation efficiency*, and (3) *insufficient pose-controllable interaction*.

Motivated by these observations, we propose **InWorld**: an **Instant-Interactive** multimodal **world** model for autonomous driving. It interactively generates sensor-level sequences in real time requiring only a single GPU, as illustrated in Figure 1. InWorld adopts a two-stage "diffusion teacher + autoregressive student" training strategy. First, a high-capacity diffusion teacher is trained in a unified LiDAR-image latent space to maximize generation quality and multimodal consistency. Then, an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

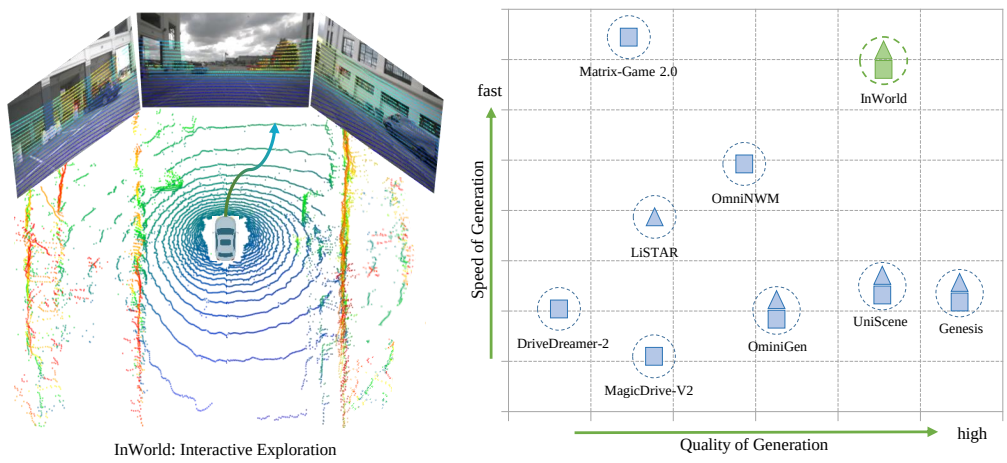


Figure 1: **Left:** InWorld enables interactive generation across views and modalities. **Right:** Generation quality versus inference efficiency; InWorld achieves the best trade-off. ▲: LiDAR generation, ■: image generation.

autoregressive student is trained via distribution matching distillation (DMD) Xu et al. (2026); He et al. (2025b); Zhen et al. (2026). The student aligns its output with the teacher's distribution along autoregressively unrolled trajectories, compressing the teacher's multi-step diffusion sampling into few-step causal sampling.

Our main contributions are as follows:

1. We introduce InWorld, a real-time, interactive multimodal world model for autonomous driving. InWorld generates synchronized RGB images and LiDAR point cloud streams at 10 Hz with precise ego-pose conditioning, providing a unified multimodal environment for end-to-end policy learning and closed-loop evaluation.
2. We propose a dedicated LiDAR VAE optimized for range image representation, and an epipolar-masked attention module. These components integrate image and LiDAR modalities into a unified latent space to enforce geometrically consistent cross-modal generation.
3. We propose a temporal decay KV caching with an autoregressive sliding window, which is integrated into a teacher–student distillation pipeline. This architecture efficiently distills a high-capacity diffusion teacher model into a lightweight autoregressive student model that preserves cross-modal consistency at 10 Hz inference.

2 Related Work

Vision-only world models Vision-only world models aim to generate visually realistic driving scene videos. DriveDreamer Wang et al. (2024); Zhao et al. (2025b) establishes a diffusion-based framework for multi-view video generation with trajectory conditioning. MagicDrive Gao et al. (2024a; 2025) further incorporates 3D geometric constraints for high-resolution long video synthesis. To address temporal consistency, Epona Zhang et al. (2025) and Panacea Wen et al. (2024) explore autoregressive diffusion and panoramic modeling, respectively. WorldLens Liang et al. (2025) provides a systematic evaluation of visual fidelity across existing generation models. These models, however, lack LiDAR simulation and few-step sampling, limiting them to offline visual synthesis.

Multimodal world models Multimodal world models have become a key technical route for end-to-end driving perception and planning. UniDriveDreamer Zhao et al. (2026) achieves single-stage joint image-LiDAR generation. UniScene Li et al. (2025) enriches geometric representation through occupancy grids. BEVWorld Zhang et al. (2024), Genesis Guo et al. (2025), and Muvo Bogdoll et al. (2023) demonstrate that unified BEV/spatiotemporal latent spaces effectively improve cross-modal consistency. For LiDAR-specific generation, LidarDM Zyrianov et al. (2024) combines

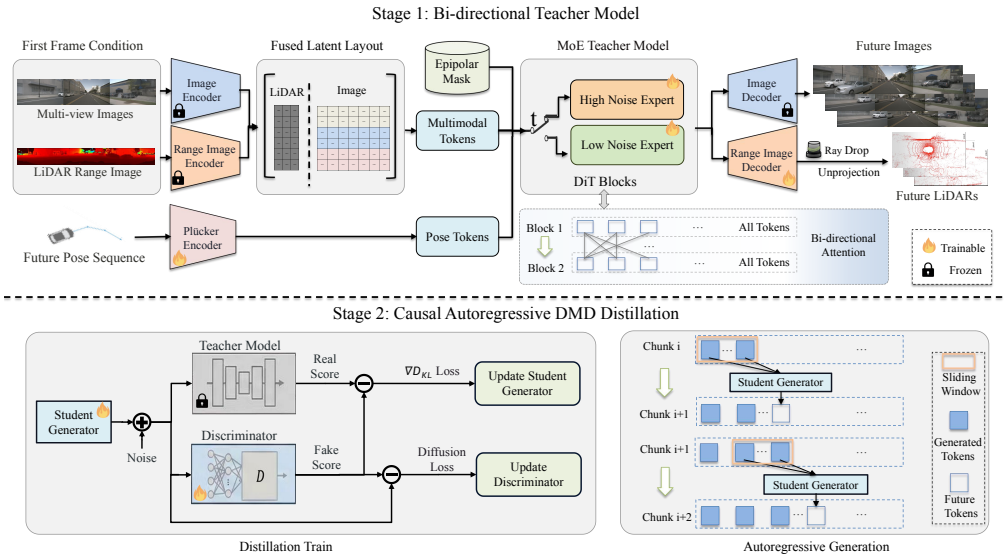


Figure 2: Overall framework of InWorld. The teacher pipeline performs multimodal image–LiDAR interaction in a unified latent space with epipolar masked attention, conditioned on the initial frame and future poses, and reconstructs future frames via dedicated decoders. DMD distillation with temporally decayed KV caching converts it into a causal autoregressive student for real-time closed-loop interaction.

latent diffusion with physical priors to produce temporally coherent point cloud sequences, while OLiDM Yan et al. (2025) employs object-scene progressive generation with semantic alignment to enhance foreground object quality. However, their reliance on multi-step diffusion or large Transformers during inference precludes instant generation.

Interactive world models Interactive world models aim to construct action-responsive virtual environments, but most target general or game scenarios rather than autonomous driving. Solaris Savva et al. (2026) proposes a multi-player video world model with viewpoint-consistent observations. RELIC Hong et al. (2025) introduces long-term spatial memory with KV caching Nguyen-Tri et al. (2025) for streaming interactive video. Matrix-Game He et al. (2025b); Wang et al. (2026) demonstrates open-source real-time interactive generation in game environments via few-step autoregressive diffusion. SoulX-LiveAct Zhen et al. (2026) targets hour-long real-time human animation with multimodal control. These methods are RGB-only and lack LiDAR, precluding their use as multimodal simulators for end-to-end driving.

3 METHOD

InWorld consists of two stages, as illustrated in Figure 2. We first train a multimodal diffusion teacher model that jointly generates multi-camera video and LiDAR in a unified latent space. This teacher model is then distilled into a lightweight autoregressive student model. Multimodal sensor data are encoded into a shared latent space via modality-aligned VAEs, then processed by a Plücker-conditioned Diffusion Transformer (DiT) with epipolar masked attention, and finally decoded into synchronized images and LiDAR via modality-specific decoders. Details of the LiDAR modeling, multimodal unified modeling and autoregressive student distillation are provided in Section 3.1, 3.2, and 3.3, respectively.

3.1 LIDAR MODELING

Range image representation and VAE encoding We adopt the range image as a structured representation of unordered, sparse point clouds. The limited vertical beam count of automotive LiDAR causes a pronounced striping pattern along the elevation axis. Following Ren et al. (2025),

we replicate each row four times consecutively along the height dimension to enhance the vertical receptive field and feature continuity of convolutional networks, yielding $I_{\text{LI}} \in \mathbb{R}^{4H\times W\times C_{\text{raw}}}$, where C_{raw} denotes the channel dimension. We then replicate channels to produce a three-channel input, enabling direct reuse of Wan2.1's image VAE architecture and pretrained weights. The VAE encoder E_{lidar} employs the same $4\times$ temporal and $8\times$ spatial compression ratios as the image, and it encodes from $\mathbb{R}^{T\times 4H\times W\times 3}$ to $\mathbb{R}^{\frac{T}{4}\times\frac{H}{2}\times\frac{W}{8}\times d}$, where d is the latent channel dimension. We initialize the LiDAR VAE from the image VAE but freeze its encoder, training only the decoder and the ray-drop head. This keeps LiDAR latents within the shared image–LiDAR latent space; joint training would let the LiDAR distribution drift and force the DiT to model two heterogeneous latent spaces.

The range image exhibits periodic boundaries along the horizontal axis. Applying standard convolutions directly introduces artifacts at these boundaries. To address this, we replace all spatial convolution operators in the VAE with horizontal circular convolutions Schubert et al. (2019): features are circularly padded and convolved along the width dimension, explicitly adapting to the annular topology of LiDAR data and eliminating representation discontinuities at the wrap-around boundary.

Ray-drop modeling Direct unprojection of simulated range images produces point clouds with uniform angular coverage. In contrast, real LiDAR scans frequently exhibit ray-drop due to physical range limits, occlusions, and low surface reflectivity. Inspired by the ray-drop modeling approach in GS-LiDAR Jiang et al. (2025), we introduce a differentiable ray-drop module at the decoder output to explicitly model whether each ray returns a measurement. Specifically, the module takes the decoded feature map I_{dec} as input and predicts a per-pixel ray return probability $p_{u,v}$ via a lightweight CNN-MLP, simulating ray loss caused by occlusion, range limits, and surface reflectance properties. During training, the module is supervised with binary ray-drop labels from the ground-truth range image using an L1 loss.

3.2 Multimodal unified modeling

Unified latent space To achieve spatiotemporal consistency for multi-view cameras and LiDAR, InWorld adopts a joint generation paradigm with a shared latent layout. Let the temporal LiDAR latent tensor be $Z_r \in \mathbb{R}^{T\times H_l\times W_l\times C}$, and the multi-view camera latent tensor (including front, front-left and front-right views) be $Z_{\text{cam}} \in \mathbb{R}^{T\times 3H_i\times W_i\times C}$. We vertically concatenate multi-view camera latent features and horizontally concatenate the transposed LiDAR features. Zero-padding is applied to unify the feature height to $3H_i$, resulting in the unified multimodal latent layout $Z_{\text{layout}} \in \mathbb{R}^{T\times(W_l+W_i)\times 3H_i\times C}$. A shared patch embedding layer projects this unified layout into token sequences that are fed into the DiT backbone for joint temporal modeling and cross-modal interaction. After denoising, the latent features of each modality are decoupled from Z_{layout} under predefined spatial rules. Each modality is then decoded independently. This follows the "latent-space concatenation — joint diffusion — disentangled decoding" pipeline of UniDriveDreamer Zhao et al. (2026), ensuring global structural consistency across modalities.

Epipolar-masked attention To further improve cross-view and cross-modal consistency, we introduce an epipolar-masked attention in the DiT self-attention modules. We use calibrated intrinsic and extrinsic parameters to back-project LiDAR rays and camera pixels into 3D space. We sample observation points within the intersection of visibility frustums. Then we test whether tokens from different sensors observe the same voxel Reading et al. (2021). A co-visibility relation sets the corresponding mask entry to 1, which enables cross-modal attention. Otherwise, the entry remains 0 and blocks interaction. Incorporated into the DiT self-attention as a structural prior, the mask encodes cross-modal token correspondences and injects 3D geometric constraints, suppressing spurious alignment while preserving multi-view geometry and LiDAR–image consistency.

Pose condition Adhering to a geometry-first paradigm He et al. (2025b), we replace text prompts with explicit geometric control. Consequently, we employ Plücker coordinates for geometric conditioning. For each pixel (u, v), the camera center $\mathbf{o}$ and ray direction $\mathbf{d}_{u,v}$ are mapped to a six-dimensional Plücker vector He et al. (2025a) as

$$\mathbf{p}_{u,v} = (\mathbf{o} \times \mathbf{d}_{u,v},\ \mathbf{d}_{u,v}) \in \mathbb{R}^6, \tag{1}$$

which parameterizes each pixel ray as a 3D line and inherently preserves multi-view projective consistency. A lightweight encoder projects $\mathbf{p}_{u,v}$ into conditioning tokens, and these tokens are

subsequently injected into the Diffusion Transformer (DiT) alongside timestamp embeddings. This architecture enables precise, pose-controllable multimodal synthesis without text conditions. Both camera and LiDAR rays are anchored to the first-frame ego pose as the reference frame, yielding an explicit 3D correspondence for every pixel and every LiDAR ray.

3.3 Real-time interactive distillation

Although the DiT-based diffusion teacher attends bidirectionally across all frames of a clip and thereby achieves high generation quality, its multi-step sampling process cannot meet the real-time requirement of closed-loop control. To address this, we introduce a distillation framework based on the Self-Forcing Huang et al. (2025) strategy, distilling the teacher's multi-step denoising capability into a few-step causal autoregressive student generator g_ϕ. Let τ denote the generation step and $\mathbf{z}_\tau$ denote the latent at step τ. The student evolves recursively as

$$\mathbf{z}_\tau = g_\phi\big(\mathbf{z}_{\tau-1},\, \mathbf{c},\, \tau\big), \quad \tau = 1, \ldots, T, \tag{2}$$

where $\mathbf{c}$ is the corresponding condition, which includes the first-frame latent and Plücker coordinates of future poses. To align the student's trajectory distribution with the teacher's, we introduce a distribution matching loss over the entire unrolled trajectory. Following the variational gradient derivation in DMD Yin et al. (2024), the parameter update direction is

$$\nabla_\phi \mathcal{L}_{\text{DMD}} = \mathbb{E}[w(t)(\hat{\epsilon}_\theta(\mathbf{z}_{\tau,t},\, \mathbf{c},\, t) - \hat{\epsilon}_\phi(\mathbf{z}_{\tau,t},\, \mathbf{c},\, t))\, \nabla_\phi\, \mathbf{z}_\tau]\,, \tag{3}$$

where t is a randomly sampled diffusion timestep, $\mathbf{z}_{\tau,t}$ is the noised version of the student output $\mathbf{z}_\tau$, and $w(t)$ is a weighting factor controlling gradient magnitude at different noise levels. $\hat{\epsilon}_\theta$ is the pretrained teacher model and remains frozen throughout, while $\hat{\epsilon}_\phi$ serves as the discriminator, trained jointly with a standard denoising diffusion loss on the noised generator outputs to estimate the score of the current student distribution. Notably, $\nabla_\phi\, \mathbf{z}_\tau$ encompasses the full gradient chain through the Self-Forcing autoregressive unrolling path.

To keep the inference cost bounded during autoregressive generation, we introduce a temporal decay KV caching mechanism. We permanently cache the first latent frame while only retaining the most recent tokens for subsequent generated frames. During training, we employ a progressive optimization paradigm that optimizes only the latest 10 latent frames in a stage-wise manner. To address the training-inference distribution mismatch, we restrict model inputs to its own historical generations, like Self-Forcing Huang et al. (2025).

3.4 Training objective

Flow matching for the diffusion teacher The core training objective follows the conditional Flow Matching framework Liu et al. (2022). Given the VAE-encoded unified multimodal latent $\mathbf{s}_0 = Z_{\text{layout}}$ and the condition $\mathbf{c}$, we construct a linear interpolation path between the data sample and a Gaussian prior $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ via:

$$\mathcal{L}_{\text{teacher}} = \mathbb{E}_{\mathbf{s}_0,\, \boldsymbol{\epsilon},\, t}\Big[\big\|\mathbf{v}_\theta((1-t)\,\mathbf{s}_0 + t\,\boldsymbol{\epsilon},\; t,\; \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{s}_0)\big\|^2\Big]\,, \tag{4}$$

where $t \sim \mathcal{U}(0, 1)$ is the diffusion timestep and $\mathbf{v}_\theta$ is the velocity field parameterized by the DiT backbone. This objective jointly optimizes cross-modal consistency within the unified latent space.

Distribution matching for the student The student is distilled from the teacher via the DMD loss Yin et al. (2024), as defined in Eq. 3.

4 Experiments

4.1 Experimental setup and baselines

Dataset and Protocol We conduct all experiments on the nuScenes Caesar et al. (2020) autonomous driving dataset, using the official training and validation splits. We adopt a sensor configuration of three cameras and one LiDAR, with each camera view set to 432×768 resolution and the LiDAR range image to 32×1088. Following common practice in driving video generation, we adopt the

short-sequence evaluation protocol, where FID and FVD are computed on fixed-length clips rather than full 241-frame (20s) scenes Gao et al. (2025; 2024b); Zhao et al. (2026); Wang et al. (2025); Gao et al. (2024a); Wang et al. (2024); Liang et al. (2025). Given the first-frame observation and future ego-pose conditions, InWorld generates an unconstrained future evolution, where other agents and background details are not required to match the ground truth trajectories frame-by-frame Feng et al. (2025).

Training and evaluation are conducted on 4 nodes with 32 NVIDIA H800 GPUs in total, whereas all inference-speed measurements are performed on a **single** H800 GPU (Sec. 4.2). The DiT model is initialized from the pretrained Wan-2.1 T2V-1.3B weights Wan et al. (2025). For the teacher model, we switch the high- and low-noise experts at a routing boundary of 947.0 following the open-source LingBot-World implementation Robbyant Team et al. (2026). During the DMD distillation stage, the generator is trained with a learning rate of 2×10^{-6}, while the discriminator uses 4.0×10^{-7}; the generator is updated once every 5 updates of the discriminator.

Baselines and evaluation metrics We compare our method with state-of-the-art existing approaches, which can be broadly categorized into three groups: camera-only, LiDAR-only, and camera+LiDAR multimodal methods. Baseline values are quoted from the original publications and should be interpreted in the context of their respective evaluation settings. We evaluate our method across five key dimensions: image/video quality, LiDAR quality, cross-modal consistency, pose control accuracy, and inference speed.

Table 1: Quantitative multimodal generation results on nuScenes. For clarity, all MMD values are scaled by 10^4.

Modality	Method	Camera Generation		LiDAR Generation		Speed
		FID ↓	FVD ↓	MMD ↓	JSD ↓	FPS ↑
Camera	MagicDrive 2024a	16.20	-	-	-	< 1
	Panacea 2024	16.96	139.0	-	-	< 1
	Drive-WM 2023	15.80	122.70	-	-	< 1
	DriveDreamer 2024	14.90	340.80	-	-	< 1
	DriveDreamer-2 2025b	11.20	55.70	-	-	< 1
	MagicDrive-V2 2025	20.91	94.84	-	-	< 1
	Epona 2025	7.5	82.8	-	-	1-5
LiDAR	LiDARVAE 2019	-	-	11.0	-	< 1
	LiDARGen 2022	-	-	19.0	0.160	< 1
	RangeLDM 2024	-	-	1.90	0.054	4.9
Camera + LiDAR	OmniGen 2025	21.01	-	2.94	0.105	< 1
	UniScene 2025	6.45	71.94	2.40	0.108	< 1
	Genesis 2025	4.24	**16.95**	-	-	< 1
	X-Drive 2024	17.37	-	**1.20**	**0.052**	-
	OmniNWM 2026	5.45	23.63	-	-	-
	InWorld (Teacher)	**3.07**	28.10	2.47	0.107	< 1
	InWorld (Student)	6.24	35.49	3.87	0.124	**10.2**

4.2 Image and LiDAR generation quality

All baseline values in Table 1 are quoted from the original publications, which report them on the nuScenes validation set. Since each table evaluates a different capability, we include a method only when it reports that capability under a comparable protocol, leaving the cell empty when the metric is unsupported or unreported; Appendix Table 5 consolidates all methods and metrics and marks these as "N/A". Table 1 first evaluates distribution-level quality on multi-view RGB videos and LiDAR. Compared with existing multimodal baselines, InWorld demonstrates competitive image quality. For LiDAR generation, the teacher attains MMD 2.47 and JSD 0.107, better than the LiDAR-only LiDARVAE and LiDARGen baselines and comparable to UniScene, indicating that the range image VAE and ray-drop module capture realistic sensor characteristics. Figure 3 shows consistent dynamic details and strong cross-modal spatial alignment.

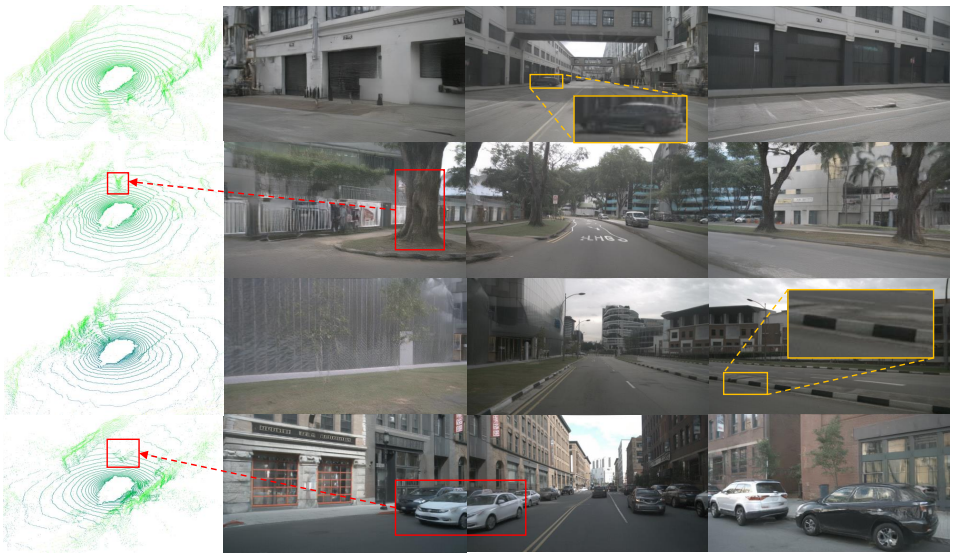


Figure 3: Qualitative results of generated LiDAR point clouds and images. Rows correspond to frames 40, 50, 60, and 80 of different generated sequences. Yellow boxes highlight zoomed-in scene structure and dynamic objects; red boxes indicate image–LiDAR alignment.

Notably, Table 1 reports both the teacher and the student: the teacher attains the best FID among all compared methods, whereas the student trades some generation quality for higher inference throughput. This gap reflects the degradation inherent to converting bidirectional diffusion into causal autoregressive generation Yin et al. (2025), which distillation only partially recovers, rather than inferior modeling capacity. Even so, the student remains superior to OmniGen and X-Drive in FID and to UniScene in FVD, while maintaining comparable generation quality and cross-modal consistency, and generating at **10 Hz (real-time)**. All inference-speed measurements use average FPS on a **single** H800 GPU, with the 32-GPU cluster used for training and evaluation-set generation only.

4.3 Cross-modal consistency evaluation

To characterize InWorld's ability in both multi-view alignment and image–LiDAR alignment, we focus on two representative consistency metrics: the View Consistency Score (VSC Swerdlow et al. (2024)) and the Depth Alignment Score (DAS Xie et al. (2024)). Both metrics rely solely on generated data and sensor calibration parameters, without requiring additional annotations.

VSC measures the consistency of keypoint correspondences in overlapping regions between adjacent camera views. For clarity, the VSC score of real images is normalized to 1.0. DAS evaluates cross-modal geometric consistency. Specifically, a monocular depth estimator Yang et al. (2024) is applied to the generated images to obtain dense depth predictions, while the generated LiDAR point clouds are projected onto the corresponding camera planes to produce sparse depth maps. The average L1 error is computed over all valid pixels between the two.

We compare InWorld with representative multi-view and multimodal generation methods that report VSC or DAS under a comparable protocol; Genesis Guo et al. (2025), for instance, reports neither and is therefore not listed. We additionally include the hybrid baseline constructed in X-Drive Xie et al. (2024), which combines MagicDrive and RangeLDM for multimodal generation (denoted as RangeLDM in the table). The quantitative results are summarized in Table 2, while qualitative comparisons are illustrated in Fig. 4. As shown, InWorld achieves the best performance on both metrics, indicating that the epipolar-masked attention and unified Plücker encoding effectively enhance multi-view and cross-modal alignment.

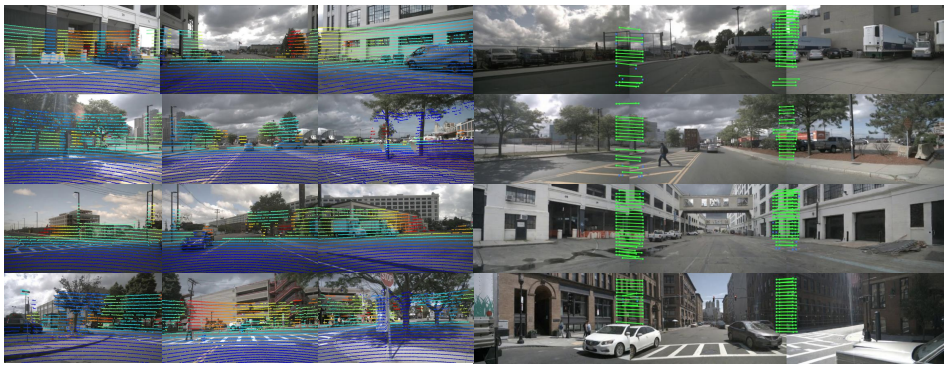


Figure 4: Visualization of cross-modal (left) and multi-view (right) consistency. Left: generated LiDAR point clouds projected onto image views, colored by depth; object boundaries and scene structure remain aligned. Right: feature points matched across overlapping regions of adjacent views, connected by line segments, showing strong geometric consistency.

Table 2: Comparison of multi-view and cross-modal consistency.

Method	VSC ↑	DAS (m) ↓
BEVGen 2024	0.439	-
MagicDrive 2024a	0.633	-
RangeLDM 2024	-	2.32
X-Drive 2024	0.675	1.69
InWorld (Teacher)	**0.852**	**1.54**
InWorld (Student)	0.829	1.61

Table 3: Pose control accuracy (R-Err / T-Err are both scaled by $\times 10^{-2}$).

Method	R-Err ↓	T-Err ↓
UniScene 2025	1.98	12.74
OmniNWM 2026	2.11	14.19
OmniNWM-Norm 2026	1.42	**5.14**
InWorld (Teacher)	**1.17**	7.38
InWorld (Student)	2.63	19.10

4.4 Trajectory control accuracy

To further evaluate the geometric accuracy under ego-pose control, we follow the evaluation protocol of OmniNWM Li et al. (2026). Given the future ground-truth poses, each method generates multi-view videos conditioned on the first-frame observation and future poses. We then apply a pretrained camera pose estimation network (same setup as OmniNWM) to the generated videos. Predicted trajectories are aligned with ground truth under a Sim(3) transformation, and we report the average rotation error (RotErr) and translation error (TransErr).

Table 3 reports the results. InWorld (Teacher) achieves the lowest reported rotation error among the listed methods, while the teacher model attains translation error close to the current best method. This advantage stems from the multimodal joint generation design: the 360° LiDAR provides dense geometric constraints that enhance rotational awareness, while the Plücker-based pose encoding enforces stronger multi-view geometric constraints, suppressing rotation errors. Although the student model exhibits approximately $2.5\times$ higher error than the teacher, it achieves over $10\times$ faster inference than the teacher with a 50% reduction in model size.

Crucially, degraded pose recovery does not imply degraded cross-view and cross-modal consistency: pose recovery measures agreement with the conditioning trajectory after Sim(3) alignment, whereas VSC and DAS measure whether the generated views and modalities are mutually consistent *with each other*. The student can thus remain internally self-consistent while deviating from the requested trajectory: from teacher to student, R-Err rises $1.17 \rightarrow 2.63$ and T-Err $7.38 \rightarrow 19.10$, while VSC ($0.852 \rightarrow 0.829$) and DAS ($1.54 \rightarrow 1.61$) change only marginally, and the student still outperforms all compared baselines on both.

Conditioning on pose rather than low-level actions does not preclude closed-loop use: an action-conditioned interface is recovered by composing InWorld with a vehicle dynamics model, while avoiding coupling sensor generation to platform-specific actuator parameters.

Table 4: Ablation results of key components, evaluated under identical settings at a common 20-epoch teacher checkpoint.

Variant	FID↓	FVD↓	VSC↑	DAS↓
Full InWorld (Teacher)	**12.59**	**32.42**	**0.830**	**2.19**
w/o MoE	13.88	34.18	0.817	2.25
w/o epipolar mask	13.61	33.89	0.771	2.41

4.5 Ablation studies

To validate the proposed components, we ablate the MoE module and the epipolar-masked attention mechanism under a unified training setup: all variants share identical initialization, data, and hyperparameters, and are evaluated at the 20-epoch teacher checkpoint with all metrics reported within the teacher architecture. Although this setup does not fully converge the models, it suffices to reveal the relative impact of each design on cross-modal geometric alignment. We report four key metrics: FID and FVD for evaluating the efficacy of the MoE strategy, and VSC and DAS for assessing the contribution of epipolar-masked attention mechanism. Table 4 reports every cell so that each component can be examined across all metrics.

The FID and FVD results demonstrate that the MoE-based teacher yields substantial improvements in both generation quality and temporal coherence. This advantage stems from the MoE architecture's noise-dependent expert routing Cheng et al. (2025); Robbyant Team et al. (2026), which assigns each denoising step to a specialized high- or low-noise expert according to the current noise level, enabling more precise modeling of multi-scale structures and temporal dynamics while operating within a constrained computational budget.

The VSC and DAS results demonstrate that incorporating epipolar-masked attention leads to measurable gains in both multi-view consistency and depth alignment accuracy. This indicates that the epipolar mask effectively guides attention toward spatially corresponding regions, suppressing interference from irrelevant cross-view features, and enforcing stronger geometric constraints during feature matching. As a result, the model establishes more accurate spatial correspondences across views. Removing either component worsens all four reported metrics at the evaluated checkpoint, supporting the contribution of both components under this setting.

5 Conclusion

We present **InWorld**, a real-time multimodal world model that generates pose-controllable, cross-modally consistent RGB images and LiDAR point clouds at 10 Hz. It integrates a range image LiDAR VAE with circular convolutions and ray-drop modeling, epipolar-masked attention for cross-modal alignment, and a teacher-student distillation pipeline that compresses multi-step diffusion into few-step autoregressive inference. Experiments on nuScenes demonstrate competitive generation quality, strong cross-modal consistency, and real-time streaming.

Our current implementation supports only front 3-view cameras. However, commercial autonomous driving systems typically deploy 7 or more cameras for full 360° coverage. This limits InWorld's efficacy in rear and side scenarios, such as lane changes and intersections. In future work, we will scale the system to a 7-camera setup by updating the epipolar masks and latent layout to maintain cross-view geometric consistency. We will also refine LiDAR modeling to better suppress noise in challenging conditions, exploring hybrid representations and physics-informed priors. Besides, our chunk-wise autoregressive paradigm suffers from latency limitations governed by VAE compression ratio and chunk size, whose trade-off with generation quality remains to be further optimized. These improvements will further enhance InWorld's practicability as an end-to-end autonomous driving simulator.

AI use statement

In this work, we used generative AI tools for retrieval and discovery, e.g., finding related work. Specifically, generative AI tools were used to assist in searching for and discovering relevant literature, and

the AI-assisted references were subsequently verified by the authors against the original publications. All citations in this paper were checked for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Detailed model architecture, training settings, data processing, and evaluation protocols are provided in Section 4.1 and in the appendix. The anonymized code and demos are available at `https://anonymous.4open.science/r/InWorld-784A`.

References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Daniel Bogdoll, Yitian Yang, Tim Joseph, Melih Yazgan, and J. Marius Zöllner. Muvo: A multimodal generative world model for autonomous driving with geometric representations. *arXiv preprint arXiv:2311.11762*, 2023.

Lucas Caccia, Herke van Hoof, Aaron Courville, and Joelle Pineau. Deep generative modeling of lidar data. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5034–5040. IEEE, 2019.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Kun Cheng, Xiao He, Lei Yu, Zhijun Tu, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Diff-moe: Diffusion transformer with time-aware and space-adaptive experts. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 10010–10024, 2025.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations (ICLR)*, 2024a.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28135–28144, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. URL `https://www.cvlibs.net/publications/Gao2024NEURIPS.pdf`.

Xiangyu Guo, Zhanqian Wu, Kaixin Xiong, Ziyang Xu, Lijun Zhou, Gangwei Xu, Shaoqing Xu, Haiyang Sun, Bing Wang, Guang Chen, et al. Genesis: Multimodal driving scene generation with spatio-temporal and cross-modal consistency. *arXiv preprint arXiv:2506.07497*, 2025.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL `https://openreview.net/forum?id=Z4evOUYrk7`.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Qianjiang Hu, Zhimin Zhang, and Wei Hu. Rangeldm: Fast realistic lidar point cloud generation. In *European Conference on Computer Vision*, pp. 115–135. Springer, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Junzhe Jiang, Chun Gu, Yurui Chen, and Li Zhang. Gs-lidar: Generating realistic lidar point clouds with panoramic gaussian splatting. *arXiv preprint arXiv:2501.13971*, 2025.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the computer vision and pattern recognition conference*, pp. 11971–11981, 2025.

Bohan Li, Zhuang Ma, Dalong Du, Baorui Peng, Zhujin Liang, Zhenqiang Liu, Xianda Guo, Zheng Zhu, Chao Ma, Yueming Jin, Xin Jin, Hao Zhao, and Wenjun Zeng. Omninwm: Omniscient driving navigation world models. In *ECCV*, pp. 187–208, 2026.

Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Wesley Yang, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, et al. Worldlens: Full-spectrum evaluations of driving world models in real world. *arXiv preprint arXiv:2512.10958*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Quan Nguyen-Tri, Mukul Ranjan, and Zhiqiang Shen. Attention is all you need for kv cache in diffusion llms. *arXiv preprint arXiv:2510.14973*, 2025.

Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Xinze Chen, Guanghong Jia, Guan Huang, and Wenjun Mei. Recondreamer-rl: Enhancing reinforcement learning via diffusion-based scene reconstruction. *arXiv preprint arXiv:2508.08170*, 2025.

Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8555–8564, 2021.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in minecraft. *arXiv preprint arXiv:2602.22208*, 2026.

Stefan Schubert, Peer Neubert, Johannes Pöschmann, and Peter Protzel. Circular convolutional neural networks for panoramic images and laser data. In *2019 IEEE intelligent vehicles symposium (IV)*, pp. 653–660. IEEE, 2019.

Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 9(4):3578–3585, 2024.

Tao Tang, Enhui Ma, Xia Zhou, Letian Wang, Tianyi Yan, Xueyang Zhang, Kun Zhan, Peng Jia, Xianpeng Lang, Jia-Wang Bian, et al. Omnigen: Unified multimodal sensor generation for autonomous driving. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9365–9374, 2025.

GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, et al. Gigaworld-0: World models as data engine to empower embodied ai. *arXiv preprint arXiv:2511.19861*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiaodong Wang, Zhirong Wu, and Peixi Peng. Longdwm: Cross-granularity distillation for building a long-term driving world model. *arXiv preprint arXiv:2506.01546*, 2025.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pp. 55–72, 2024.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *arXiv preprint arXiv:2311.17918*, 2023.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024.

Yichen Xie, Chenfeng Xu, Chensheng Peng, Shuqi Zhao, Nhat Ho, Alexander T Pham, Mingyu Ding, Masayoshi Tomizuka, and Wei Zhan. X-drive: Cross-modality consistent multi-sensor data synthesis for driving scenarios. *arXiv preprint arXiv:2411.01123*, 2024.

Tongda Xu, Mingwei He, Shady Abu-Hussein, Jose Miguel Hernandez-Lobato, Chunhang Zheng, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction fid predictive of diffusion generation fid. *arXiv preprint arXiv:2603.05630*, 2026.

Tianyi Yan, Junbo Yin, Xianpeng Lang, Ruigang Yang, Cheng-Zhong Xu, and Jianbing Shen. Olidm: Object-aware lidar diffusion models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9121–9129, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27220–27230, 2025.

Yumeng Zhang, Shi Gong, Kaixin Xiong, Xiaoqing Ye, Xiaofan Li, Xiao Tan, Fan Wang, Jizhou Huang, Hua Wu, and Haifeng Wang. Bevworld: A multimodal world simulator for autonomous driving via scene-level bev latents. *arXiv preprint arXiv:2407.05679*, 2024.

Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, et al. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In *Proceedings of the computer vision and pattern recognition conference*, pp. 12015–12026, 2025a.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 10412–10420, 2025b.

Guosheng Zhao, Yaozeng Wang, Xiaofeng Wang, Zheng Zhu, Tingdong Yu, Guan Huang, Yongchen Zai, Ji Jiao, Changliang Xue, Xiaole Wang, et al. Unidrivedreamer: A single-stage multimodal world model for autonomous driving. *arXiv preprint arXiv:2602.02002*, 2026.

Dingcheng Zhen, Xu Zheng, Ruixin Zhang, Zhiqi Jiang, Yichao Yan, Ming Tao, and Shunshun Yin. Soulx-liveact: Towards hour-scale real-time human animation with neighbor forcing and convkv memory. *arXiv preprint arXiv:2603.11746*, 2026.

Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point clouds. In *European Conference on Computer Vision*, pp. 17–35. Springer, 2022.

Vlas Zyrianov, Henry Che, Zhijian Liu, and Shenlong Wang. Lidardm: Generative lidar simulation in a generated world. *arXiv preprint arXiv:2404.02903*, 2024.

A Implementation of epipolar self-attention

Our multi-sensor system consists of one LiDAR and three cameras. To establish explicit geometric correspondences across modalities and views, we adopt individual 3D rays as the fundamental constraint unit. For any point on the observation plane of a source sensor (e.g., an image pixel or a LiDAR range image grid cell), a unique ray is defined in 3D space.

We sample N discrete points along each ray and project them onto the observation plane of a target sensor. This yields N explicit geometric constraints, indicating that any physical point lying on the ray must be observed at locations consistent with these projected coordinates in the target sensor. We inject these constraints into the self-attention mechanism of the diffusion process via an epipolar mask, thereby enabling information aggregation exclusively within theoretically co-visible regions.

This formulation can be regarded as a generalization of classical two-view epipolar geometry to heterogeneous sensors, providing a physically grounded prior for cross-modal feature alignment.

A.1 Epipolar constraints across modalities

Let $P_x(p_i, p_j)$ denote a grid point on the observation plane of a source sensor A (pixel coordinate for cameras or range image index for LiDAR). Using the intrinsic and extrinsic parameters $(\mathbf{K}_A, \mathbf{R}_A, \mathbf{t}_A)$, we construct the corresponding ray in the world coordinate system.

For camera modalities, we back-project via $\mathbf{K}_A^{-1}$ to the normalized image plane; for LiDAR, the ray direction is determined by the azimuth and elevation angles associated with the grid index. The unified ray equation is

$$\mathbf{P}(\lambda) = \mathbf{C}_A + \lambda \mathbf{d}_A, \quad \lambda > 0, \tag{5}$$

where $\mathbf{C}_A = -\mathbf{R}_A^\top \mathbf{t}_A$ is the ray origin and $\mathbf{d}_A$ is the unit direction vector in world coordinates.

We sample $N = 1500$ points along the ray within a valid depth range $[\lambda_{\min}, \lambda_{\max}]$. To better capture near-field parallax and far-field sparsity, the sampling follows an exponential spacing strategy:

$$\lambda_m = \lambda_{\min} \cdot \exp\left(\frac{m-1}{N-1} \ln \frac{\lambda_{\max}}{\lambda_{\min}}\right), \quad m = 1, \ldots, N. \tag{6}$$

The sampled points are denoted as $\{\mathbf{P}^{(m)}\}_{m=1}^{N}$.

Each point is transformed into the local coordinate system of a target sensor B: $\mathbf{P}_B^{(m)} = \mathbf{R}_B\mathbf{P}^{(m)} + \mathbf{t}_B$, followed by modality-specific projection.

For cameras, projection follows the pinhole model

$$P_y^{(m)} = \left(\frac{[\mathbf{K}_B\mathbf{P}_B^{(m)}]_x}{[\mathbf{K}_B\mathbf{P}_B^{(m)}]_z}, \frac{[\mathbf{K}_B\mathbf{P}_B^{(m)}]_y}{[\mathbf{K}_B\mathbf{P}_B^{(m)}]_z} \right), \tag{7}$$

subject to $[\mathbf{K}_B\mathbf{P}_B^{(m)}]_z > 0$ and within the field of view.

For LiDAR, $\mathbf{P}_B^{(m)}$ is converted to spherical coordinates, and discretized into range image indices based on angular resolution.

Thus, each source point P_x and its projections $\{P_y^{(m)}\}$ define a set of cross-modal geometric constraints. This enforces that any physical point on the ray must lie within the neighborhood of these projected locations. These constraints are later used to construct binary attention masks for physically consistent feature routing.

A.2 Attention mask construction and integration

Let the total token length be N_{token}. We construct a binary mask $\mathbf{M} \in \{0,1\}^{N_{\text{token}} \times N_{\text{token}}}$ as follows.

We initialize $\mathbf{M}$ to zeros, blocking all cross-view and cross-modal interactions by default. For each token pair across sensors, we project sampled ray points and map them into discrete token indices $\mathcal{I}_{\text{proj}}$ based on the patch size P. All indices satisfying the epipolar co-visibility constraint are assigned $\mathbf{M}_{i,j} = 1$.

Due to the symmetry of physical visibility, we enforce $\mathbf{M} \leftarrow \max(\mathbf{M}, \mathbf{M}^\top)$ to ensure bidirectional consistency.

The mask is injected into the DiT attention as an additive bias as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathcal{A}(\mathbf{M}) \right)\mathbf{V}, \tag{8}$$

where $\mathcal{A}(\mathbf{M})_{i,j} = 0$ if $\mathbf{M}_{i,j} = 1$, and $-\infty$ otherwise.

This design enforces hard geometric constraints by completely suppressing gradients of invalid attention connections during backpropagation.

B More implementation details about training and model

B.1 LiDAR range image representation

Given a raw point cloud $\mathcal{P} = \{p_i = (x_i, y_i, z_i)\}_{i=1}^{N}$, we first compute the spherical coordinates of each point:

$$r_i = \sqrt{x_i^2 + y_i^2 + z_i^2} \tag{9}$$

$$\theta_i = \text{atan2}(y_i, x_i) \in [-\pi, \pi] \tag{10}$$

$$\phi_i = \arcsin\left(\frac{z_i}{r_i}\right) \in [-\pi/2, \pi/2] \tag{11}$$

and project each point onto a 2D grid $I(u, v)$ based on (θ_i, ϕ_i), populating channels with normalized range.

B.2 Model architecture and input representation

The backbone is built upon the Wan2.1 1.3B diffusion transformer (DiT), extended for multimodal generation. The original DiT contains 30 self-attention layers for modeling spatial and temporal dependencies in image modalities.

To enhance cross-modal alignment, we append 4 additional self-attention layers equipped with epipolar masks at the end of the network. These layers are interleaved with the last 4 original layers in an alternating schedule, enabling progressive fusion of image and LiDAR features across scales. The choice of 4 layers is a trade-off between cost and performance.

For input preprocessing, each image has a resolution of 432×768. Multi-frame images are concatenated vertically. The LiDAR range image is vertically flipped, repeated 4 times along height, transposed, and concatenated to the left of the image, forming a multimodal tensor of size 1296×896. This tensor is partitioned into patches using a sliding window with stride 8×8 and receptive field 2×2, yielding 4536 tokens. To improve training stability, all model parameters are maintained with exponential moving average (EMA) with decay factor 0.99.

B.3 Training strategy and optimization

DiT training We train the diffusion backbone using `bfloat16` mixed precision. Training is conducted on 32 NVIDIA H800 GPUs using FSDP, with a global batch size of 32. The learning rate follows a cosine decay schedule with linear warm-up, initialized at 1.0×10^{-5}.

Training proceeds in two stages. In stage one, only the original 30 layers are trained for 50 epochs to learn unimodal spatial and temporal priors. In stage two, the 4 epipolar-masked layers are inserted with zero initialization, preserving pretrained distributions and preventing representation collapse.

We adopt timestep rescaling with shift=10.0 to bias sampling toward higher noise levels, strengthening structural learning signals.

LiDAR VAE training Due to the numerical range of normalized LiDAR data ($[-1, 1]$), `bfloat16` may introduce quantization artifacts. Therefore, the VAE is trained entirely in `float32` precision.

We use AdamW optimizer with learning rate 2.0×10^{-5}. The loss is

$$\mathcal{L}_{\text{VAE}} = 1.0 \cdot \mathcal{L}_{\text{MSE}} + 0.3 \cdot \mathcal{L}_{\text{LPIPS}}, \tag{12}$$

where LPIPS is computed using pretrained VGG-16 features.

The VAE uses an $8\times$ downsampling architecture and is trained for 200 epochs until convergence.

C Additional evaluation details

Temporal sampling protocol Temporal metrics such as FVD are highly sensitive to clip length and sampling strategy Liang et al. (2025). Prior works (e.g., DriveDreamer, MagicDrive) adopt fixed-length short clips generated independently. We strictly follow this protocol: all evaluations are conducted on independently generated short clips, without autoregressive rollout followed by segmentation.

These two paradigms differ fundamentally: (1) Independent clip generation ensures i.i.d. samples aligned with training distribution; (2) Rolling long-sequence sampling introduces temporal error accumulation and distribution drift, leading to biased metrics. Thus, our evaluation strictly adheres to the fair comparison protocol for fair benchmarking.

C.1 Unified comparison across methods and metrics

Each table in the main paper evaluates a different capability, so the set of compared methods differs from table to table. To provide a single consolidated view, Table 5 collects every method and every metric discussed in this work. A cell is marked "N/A" when the method does not support or does not report the corresponding metric; for instance, camera-only generators produce no LiDAR and therefore have no LiDAR-quality or cross-modal score. All baseline values are quoted from the original publications. The UniScene entry corresponds to its generated-occupancy configuration, i.e., the setting that does not rely on ground-truth occupancy as input.

Table 5: Unified comparison of all methods and all metrics discussed in this work. "N/A" indicates that the method does not support or does not report the corresponding metric. FID and FVD are reported for multi-view video generation, whereas MMD and JSD are reported for LiDAR point clouds. The UniScene entry corresponds to its generated-occupancy configuration. The modality of OmniNWM is camera with metric depth and 3D occupancy (C+D+O) rather than LiDAR. The MagicDrive + RangeLDM row is the hybrid cross-modal baseline constructed in X-Drive 2024, referred to as "RangeLDM" in Table 2.

Method	Modality	FID ↓	FVD ↓	MMD ↓	JSD ↓	FPS ↑	VSC ↑	DAS ↓
Camera-only								
MagicDrive 2024a	C	16.20	N/A	N/A	N/A	< 1	0.633	N/A
Panacea 2024	C	16.96	139.00	N/A	N/A	< 1	N/A	N/A
Drive-WM 2023	C	15.80	122.70	N/A	N/A	< 1	N/A	N/A
DriveDreamer 2024	C	14.90	340.80	N/A	N/A	< 1	N/A	N/A
DriveDreamer-2 2025b	C	11.20	55.70	N/A	N/A	< 1	N/A	N/A
MagicDrive-V2 2025	C	20.91	94.84	N/A	N/A	< 1	N/A	N/A
Epona 2025	C	7.50	82.80	N/A	N/A	1–5	N/A	N/A
BEVGen 2024	C	N/A	N/A	N/A	N/A	N/A	0.439	N/A
LiDAR-only								
LiDAR VAE 2019	L	N/A	N/A	11.0	N/A	< 1	N/A	N/A
LiDARGen 2022	L	N/A	N/A	19.0	0.160	< 1	N/A	N/A
RangeLDM 2024	L	N/A	N/A	1.90	0.054	4.9	N/A	N/A
Camera + LiDAR								
OmniGen 2025	C+L	21.01	N/A	2.94	0.105	< 1	N/A	N/A
UniScene 2025	C+L	6.45	71.94	2.40	0.108	< 1	N/A	N/A
Genesis 2025	C+L	4.24	**16.95**	N/A	N/A	< 1	N/A	N/A
MagicDrive + RangeLDM 2024	C+L	N/A	N/A	N/A	N/A	N/A	N/A	2.32
X-Drive 2024	C+L	17.37	N/A	**1.20**	**0.052**	N/A	0.675	1.69
OmniNWM 2026	C+D+O	5.45	23.63	N/A	N/A	N/A	N/A	N/A
InWorld (Teacher)	C+L	**3.07**	28.10	2.47	0.107	< 1	**0.852**	**1.54**
InWorld (Student)	C+L	6.24	35.49	3.87	0.124	**10.2**	0.829	1.61

D More experimental results

D.1 LiDAR VAE performance evaluation

The quantitative results in Table 6 indicate that our LiDAR VAE achieves competitive reconstruction quality compared to prior methods. In particular, it attains the best Chamfer Distance while maintaining a high F-Score, demonstrating that the learned latent representation is sufficiently accurate for downstream multimodal generation and preserves essential geometric structures. Removing explicit ray-return modeling degrades Chamfer Distance from 0.136 to 0.167 and F-Score from 0.864 to 0.750, while circular convolution, whose effect is confined to the horizontal seam of the range image, improves boundary-region Chamfer Distance from 0.162 to 0.155 and F-Score from 0.779 to 0.787 when evaluated on the leftmost and rightmost 50 columns only.

Table 6: Comparison of LiDAR VAE performance across different methods.

Method	Chamfer Distance ↓	F-Score ↑
OmniGen 2025	0.793	0.742
UniDriveDreamer 2026	0.154	**0.900**
Ours	**0.136**	0.864
w/o ray-drop	0.167	0.750

D.2 More image and LIDAR generation results

We provide additional image and LIDAR generation results at the anonymized repository: `https://anonymous.4open.science/r/InWorld-784A`.

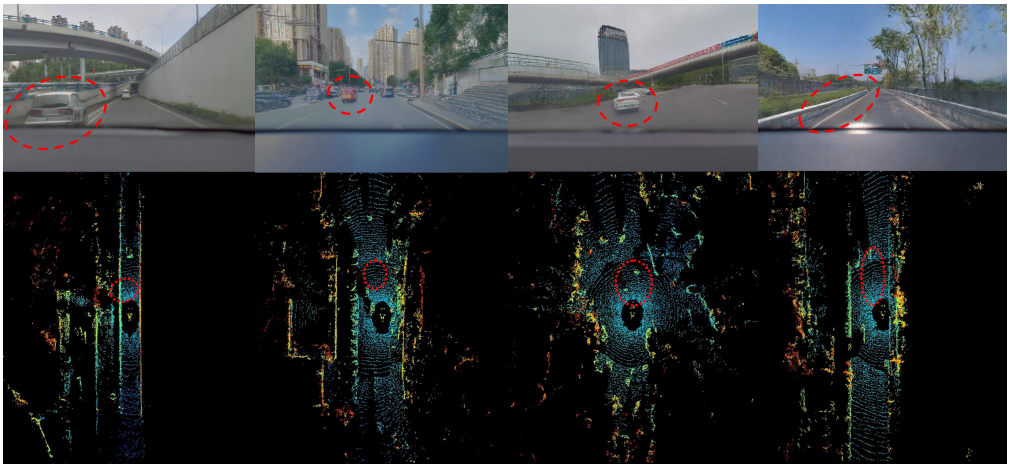


Figure 5: Failure cases of voxel-based cross-modal generation on real-world data. Top row: generated RGB images; bottom row: corresponding LiDAR point clouds in bird's-eye view. Columns 1–2: vehicles are generated in images but missing in LiDAR; Column 3: two vehicles appear in LiDAR while only one is present in the image; Column 4: spatial misalignment of roadside barriers between image and LiDAR.

E Failure cases and analysis

E.1 Voxel-based LiDAR generation

Before adopting the range image representation, we first explored a voxel-based generation paradigm (with architecture adapted from OpenDWM). When trained independently, the voxel-based DiT model can produce temporally and spatially coherent 3D point clouds. However, during joint optimization with the image DiT, stable cross-modal correspondence is difficult to establish, regardless of whether global self-attention or bidirectional cross-attention is used. As shown in Figure 5, the generated LiDAR and images exhibit significant misalignment in both semantic instances and spatial layout. Notably, introducing the epipolar mask constraint does not alleviate this issue.

We attribute this limitation to the severe representational heterogeneity between voxel and image latent spaces. Voxel features encode discrete 3D occupancy and explicit height information, whereas image features capture dense 2D photometric and perspective semantics. This discrepancy in data manifolds and inductive biases prevents attention mechanisms from effectively aligning cross-modal features, leading to unstable joint training and poor correspondence. This observation motivates our transition to the range image representation, which projects LiDAR into a 2D structured domain compatible with image tokens, ultimately enabling stable and consistent cross-modal generation.

E.2 Choice of LiDAR VAE design

In the early stage of building the multimodal framework, we experimented with several independent VAE architectures for LiDAR point cloud compression and reconstruction. However, we found that heterogeneous VAE designs induce a significant semantic gap and distribution mismatch between image and LiDAR latent representations, which slow down convergence during joint training.

We therefore adopt the Wan2.1 VAE architecture with minimal modifications. Specifically, we keep the image encoder parameters frozen and only optimize the decoder for LiDAR data. This design enforces a shared latent prior between image and LiDAR modalities, effectively eliminating

cross-modal distribution discrepancies and ensuring strong alignment in both semantic and statistical properties.