Doe-2: 3D Representation World Model for Unified Driving Scene Forecasting

Dong Zhuo[1,*], Wenzhao Zheng[1,*,†,✉], Sicheng Zuo[1], Yuanhui Huang[1]
Siming Yan[2], Lu Hou[2], Jie Zhou[1], and Jiwen Lu[1]

[1]Tsinghua University, China [2]Yinwang Intelligent Technology Co. Ltd., China
Page: `https://wzzheng.net/Doe-2/`
Code: `https://github.com/paryi555/Doe-2`

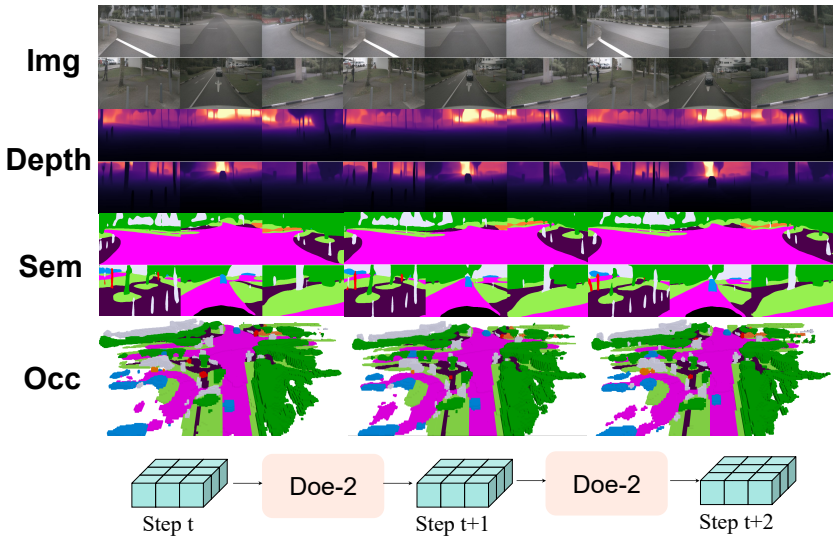

Fig. 1: Operating in a unified 3D latent representation space, Doe-2 autoregressively forecasts future scene evolutions conditioned on ego-actions. Unlike existing world models that focus solely on appearance, our native 3D world model simultaneously decodes multi-view images, depth maps, semantics, and 3D occupancy in a single pass.

Abstract. Forecasting future scene evolutions is crucial for autonomous driving safety. Most existing driving world models focus on appearance generation and fail to model geometric and semantic evolutions. In this paper, we propose a native 3D world model, **Doe-2**, to forecast the comprehensive scene evolutions in a unified 3D latent representation space. This representation contains both appearance, geometry, and semantic information and can be decoded into multi-view RGB, depth, semantics, and 3D occupancy. For efficiency, we adopt an autoregressive transformer architecture to generate all latents for the next scene in one pass and then use a lightweight diffusion head to refine them for better fidelity. Operating on the compact 3D latent space, our model supports efficient future rollouts conditioned on different actions. Extensive experiments on the nuScenes dataset show the ability of **Doe-2** to efficiently and comprehensively forecast future driving scenes. Despite being trained from scratch

*Equal contributions. †Project leader. ✉Corresponding Author.

with only 700 scenes, **Doe-2** already shows high-quality scene generation and good generalization with action control ability.

Keywords: Autonomous Driving · World Model · 3D Scene Forecasting · Unified Representation

1 Introduction

Forecasting future scene evolution is crucial for autonomous driving safety [26, 32, 49, 69, 70]. Recently, world models have emerged as a promising paradigm for simulating this evolution by predicting subsequent states based on historical context and ego-actions [13, 15, 29, 30, 67, 72]. By enabling high-fidelity simulation and precise prediction of complex environments, they have demonstrated immense potential in advancing closed-loop autonomous driving simulation as well as holistic scene perception and understanding. To build an effective autonomous driving world model, two core pillars are indispensable: a holistic 3D scene representation that accurately captures the physical environment, and a generative architecture that ensures both high-quality synthesis and agile control responsiveness. Yet, existing methods struggle to satisfy both requirements.

Representatively, current world models face a severe trade-off, as illustrated in Fig. 2. 2D-centric approaches [8, 9, 15, 48, 69] focus on appearance generation, inherently lacking geometric awareness and holistic 3D scene perception, which makes them struggle with multi-view consistency. Conversely, explicit 3D methods [4, 11, 61, 63, 67] relying on point clouds or occupancy grids as inputs suffer from expensive annotation costs and computational inefficiency. Generatively, these models are trapped in an architectural dilemma. Autoregressive (AR) models excel at capturing sensitive temporal dynamics and respond agilely to action conditions, making them ideal for sequential decision-making [40,50,67]. However, they often yield blurry or low-fidelity visual results [2]. On the other hand, while Diffusion Transformers (DiT) are capable of high-fidelity synthesis, directly generating high-dimensional scene representations exposes them to sluggish iterative denoising and severe degradation in generation quality [10, 27, 66]. Furthermore, conventional multimodal world models often rely on cumbersome multi-stage pipelines or fragmented generative pathways [24, 26, 56], leading to sub-optimal feature alignment and compounding errors.

To resolve these challenges, we propose Doe-2, a native 3D representation world model for unified driving scene forecasting. Relying solely on multi-view images, Doe-2 elegantly bridges the gap between holistic 3D representation and efficient, high-quality generation. Specifically, we build upon a unified 3D scene tokenizer, DriveTok [73], and introduce a feature compressor inspired by FAE [10] to encode multi-view semantics and dense geometry into a highly compact 3D latent space. Building on this compact space, an action-conditioned autoregressive (AR) architecture captures spatio-temporal dynamics and models the structural layout of future latents, ensuring agile control. Crucially, recent studies [10,27,66] reveal that autoregressive modeling and low-dimensional latent spaces can significantly simplify diffusion tasks. Inspired by this, we append a lightweight

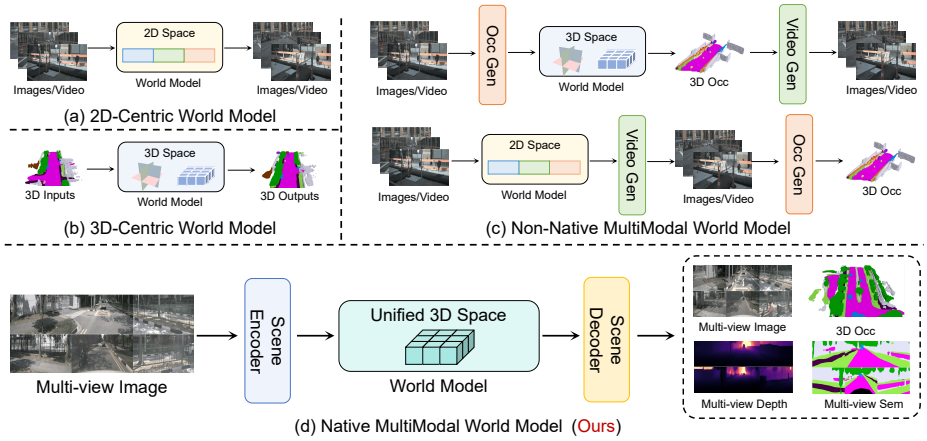

Fig. 2: Comparison of different world model paradigms. (a) 2D-Centric World Models operate entirely in 2D space, lacking geometric awareness and multi-view consistency.(b) 3D-Centric World Models rely on 3D inputs, which are computationally expensive and annotation-heavy. (c) Non-Native MultiModal World Models employ a cascaded architecture (e.g., relying on occupancy or video generators), which introduces error accumulation. (d) Our Doe-2 directly encodes multi-view images into a unified 3D space, decoding multiple modalities simultaneously in an end-to-end manner.

3-layer DiT head for high-fidelity refinement. With global dependencies already resolved by the AR model, the DiT focuses solely on injecting high-frequency details, bypassing the computational overhead of full-scale diffusion.

In Doe-2, the predicted future latents are recursively fed back for stable extrapolation as shown in Fig. 1, yielding highly controllable and temporally consistent future rollouts across diverse driving scenarios. Trained entirely from scratch on the nuScenes [3] dataset, Doe-2 decodes these refined latents to simultaneously generate multi-view images, 3D occupancy, depth, and semantics in a single pass. By obviating the need for cascaded pipelines in Fig. 2 (c), this native unified approach, as shown in Fig. 2 (d), inherently mitigates compounding errors and significantly boosts inference efficiency. Extensive experiments demonstrate that Doe-2 achieves impressive results in controllable visual synthesis and accurate 3D geometric forecasting, setting a new standard for scalable closed-loop simulation and ultimately paving the way for safer autonomous driving systems.

2 Related Work

2D-Centric World Models for Autonomous Driving. Early world models primarily operate in the 2D image or video space. DriveDreamer [48] builds a real-world diffusion-based driving world model with traffic conditioning and action-driven future prediction. Building on scalable sequence modeling, GAIA-1 [15] reformulates world modeling as autoregressive latent token dynamics coupled with a decoder for multimodal control. Moving toward geometric control, MagicDrive [8] introduces 3D-aware conditioning and cross-view attention for multi-camera consistency, addressing limitations of 2D layout guidance. Vista [9]

enhances fidelity and controllability by injecting dynamic priors and designing dynamics aware losses with unified multi-level action control. GEM [14] generalizes to multimodal world modeling with fine-grained control over ego-motion, object dynamics, and human poses. Although these models achieve impressive visual fidelity, their focus on appearance generation, operating exclusively within the 2D image domain restricts their geometric awareness.

3D-Centric World Models for Autonomous Driving. To address the lack of spatial awareness, subsequent works explicitly model the world in structured 3D spaces. OccWorld [67] and DFIT-OccWorld [61] cast world modeling as autoregressive prediction in 3D semantic occupancy grids, shifting from pixel-space generation to voxel-level spatial dynamics. DrivingSphere [56] builds a composed 4D occupancy world for closed-loop simulation. In contrast, BEVWorld [63] adopts a unified BEV latent space to fuse multi-view images and LiDAR, performing action-conditioned diffusion directly in bird's-eye-view representations. DriveWorld [40] further frames world modeling as 4D spatio-temporal representation learning, combining dynamic memory states with 3D occupancy prediction for downstream tasks. GeoDrive [4] explicitly integrates 3D point conditions into driving world models, leveraging 3D structural priors derived from monocular inputs to enhance spatial understanding and action-conditioned generation. DriveDreamer4D [64] enhances 4D Gaussian scene representations by injecting world-model priors to improve novel-trajectory rendering. While explicit 3D representations provide the accurate spatial awareness crucial for planning, operating directly in these high-dimensional spaces is computationally prohibitive and heavily reliant on costly 3D inputs.

Multimodal World Models for Autonomous Driving. Recognizing the necessity of holistic scene understanding, recent efforts move beyond single-modality prediction toward joint generation. UniScene [24] adopts an occupancy-centric framework to unify RGB, LiDAR, and 3D occupancy in a shared representation space. Genesis [13] synthesizes multi-view driving videos and LiDAR sequences within a unified two-stage framework, enhancing cross-modal consistency and controllability through structured semantic supervision. OmniNWM [26] generates panoramic RGB, semantics, depth, and 3D occupancy while integrating action conditioning and intrinsic occupancy-grounded rewards. The HERMES series [71, 72] has taken a significant step toward a unified world model for autonomous driving that integrates understanding and generation, while also exploring their collaborative nature. Despite their comprehensive coverage, many of these multimodal frameworks rely on fragmented, multi-stage generative pipelines, which inevitably lead to feature misalignment and cascading errors.

3 Proposed Approach

3.1 Overall Architecture

As shown in Fig. 3, we decompose the task of controllable driving scene generation into three functional stages. First, we leverage the pretrained DriveTok [73]

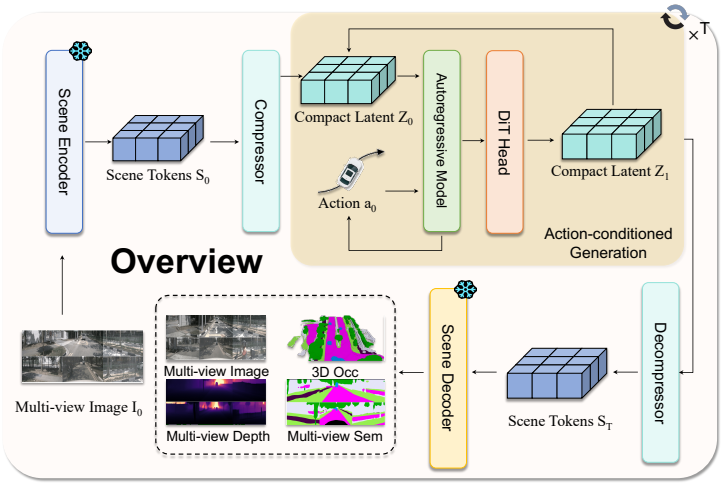

Fig. 3: Overall architecture of our proposed framework. The pipeline is structured into three key stages. Initially, multi-view images are encoded into a unified 3D representation via a frozen Scene Encoder. Next, a Compressor maps these features into a highly efficient compact latent space. Within this space, an action-conditioned generation module predicts future states. Finally, a Decompressor and Scene Decoder simultaneously output diverse 2D and 3D modalities in a single pass.

as a scene encoder to extract a unified 3D scene representation $\mathbf{S}_t$ from multi-view inputs $\mathbf{I}_t$, and then decode the predicted future representation $\hat{\mathbf{S}}_{t+1}$ into multimodal outputs $\hat{\mathbf{Y}}_{t+1}$ (Sec. 3.2). Second, to alleviate computational overhead and facilitate high-quality generation, a paired feature compressor and decompressor are introduced to encode $\mathbf{S}_t$ into a compact latent space $\mathbf{Z}_t$ and subsequently decompress the predicted latents $\hat{\mathbf{Z}}_{t+1}$ (Sec. 3.3). Finally, by operating in this efficient latent space, our action-conditioned generative module is able to predict the controllable future latents $\hat{\mathbf{Z}}_{t+1}$ (Sec. 3.4). Notably, our final decoding stage generates multi-view images, 3D occupancy, depth, and semantics simultaneously in a single pass, preventing compounding errors and latency.

3.2 Unified 3D Scene Representation

Recent 2D-centric world models have demonstrated remarkable generative capabilities. However, these works lack an explicit 3D scene representation and process multi-view inputs solely in the 2D image/video domain, which limits their geometric reasoning capabilities. Furthermore, tokenizing high-resolution multi-view inputs independently results in a massive number of visual tokens, imposing prohibitive computational burdens.

To overcome these limitations, we formulate our world model upon a native unified 3D scene representation. Specifically, we leverage the pretrained DriveTok [73] as our scene tokenizer. Given the multi-view images $\mathbf{I}_t \in \mathbb{R}^{N\times 3\times H\times W}$ from N cameras at time step t, the scene encoder of DriveTok, denoted as $\mathcal{E}_{scene}$, encodes them into a set of 3D scene tokens $\mathbf{S}_t \in \mathbb{R}^{C\times H_s\times W_s}$:

$$\mathbf{S}_t = \mathcal{E}_{scene}(\mathbf{I}_t). \tag{1}$$

By internally utilizing 2D foundation models and 3D deformable cross-attention, $\mathcal{E}_{scene}$ ensures that the extracted $\mathbf{S}_t$ integrates textural, semantic, and geometric information. More importantly, these scene tokens are inherently agnostic to camera layout and image resolution, providing a unified and efficient representation for downstream temporal modeling.

Symmetrically, during the decoding phase, $\mathcal{D}_{scene}$ is employed to decode multimodal outputs from the predicted future scene representation $\hat{\mathbf{S}}_{t+1}$:

$$\hat{\mathbf{Y}}_{t+1} = \mathcal{D}_{scene}(\hat{\mathbf{S}}_{t+1}), \tag{2}$$

where $\hat{\mathbf{Y}}_{t+1}$ encompasses multi-view RGB images, depth maps, 2D semantics, and 3D occupancy. By anchoring both the encoding and decoding processes in this unified 3D token space, our framework naturally ensures cross-view geometric consistency and robust 3D spatial awareness, laying a solid foundation for the subsequent autoregressive generation.

3.3 3D Scene Compression

While the unified 3D scene representation $\mathbf{S}_t$ extracted by DriveTok contains comprehensive geometric and semantic information, directly applying Diffusion Transformers (DiTs) to such high-dimensional 3D representations leads to sluggish iterative denoising and degradation in generation quality. To bridge this gap, we introduce a feature compression module inspired by FAE [10].

Specifically, a lightweight compressor $\mathcal{E}_{comp}$ compresses $\mathbf{S}_t$ into a low-dimensional latent space $\mathbf{Z}_t \in \mathbb{R}^{D \times H_s \times W_s}$, where the channel dimension is substantially reduced ($D \ll C$). To prevent overfitting to the relatively simple reconstruction objective and retain critical pre-trained semantics and geometry, $\mathcal{E}_{comp}$ is composed of only a single self-attention layer and a linear projection. The compressor outputs the mean $\boldsymbol{\mu}_t \in \mathbb{R}^{D \times H_s \times W_s}$ and logarithmic variance $\log \boldsymbol{\sigma}_t^2 \in \mathbb{R}^{D \times H_s \times W_s}$:

$$\boldsymbol{\mu}_t, \log \boldsymbol{\sigma}_t^2 = \mathcal{E}_{comp}(\mathbf{S}_t). \tag{3}$$

During training, we sample the latent representation $\mathbf{Z}_t$ via the reparameterization:

$$\mathbf{Z}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}_t, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

Moreover, we adopt an asymmetric generative training strategy. The stochastic sampled latent $\mathbf{Z}_t$ serves as the input for autoregressive modeling to enhance robustness, while the deterministic mean $\boldsymbol{\mu}_t$ acts as the stable target for the generative module, effectively simplifying the objective of generation.

Finally, a decompressor $\mathcal{D}_{comp}$ reconstructs the high-dimensional scene tokens from the predicted future latents $\hat{\mathbf{Z}}_{t+1}$:

$$\hat{\mathbf{S}}_{t+1} = \mathcal{D}_{comp}(\hat{\mathbf{Z}}_{t+1}), \tag{5}$$

where the $\mathcal{D}_{comp}$ employs a 6-layer transformer architecture equipped with components like RoPE, RMSNorm, and SwiGLU. This compression paradigm drastically reduces computational overhead, transforming a prohibitive 3D diffusion task into an efficient latent-space generation process.

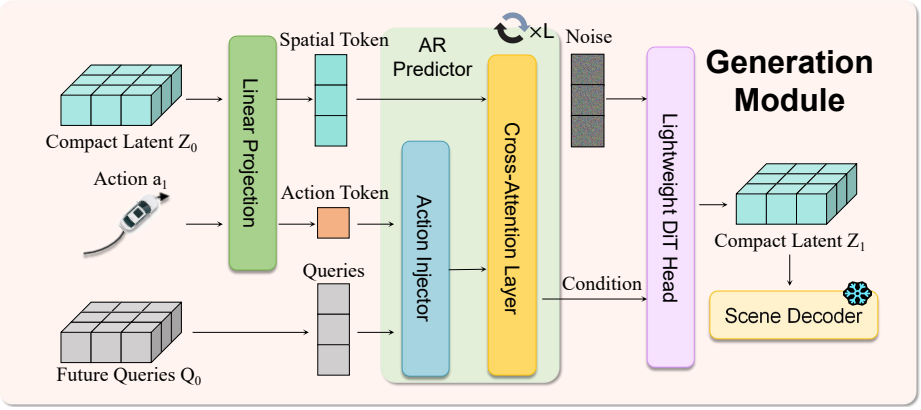

Fig. 4: Forecasting architecture of Doe-2. Conditioned on ego-action, the AR Predictor processes future queries to capture spatio-temporal dynamics. A lightweight DiT head then refines these predictions into a high-fidelity, compact 3D latent, which is finally decoded into comprehensive multi-modal scene outputs.

3.4 Action-Conditioned Generative Model

Prior works [2, 27] have shown that AR models excel at action-conditioned temporal dynamics but lack visual fidelity, whereas DiTs generate high-quality details but struggle with sluggish inference. To harness their complementary strengths, we propose a hybrid paradigm as shown in Fig. 4. By employing an AR model to resolve global spatio-temporal structures and forecast coarse latents, the diffusion target is significantly simplified. This enables a lightweight 3-layer DiT head to bypass full-scale diffusion overhead, focusing exclusively on refining these coarse latents.

Autoregressive Prior Prediction. Given historical latents $\mathbf{Z}_{\leq t}$ and ego-actions $\mathbf{a}_{\leq t}$, the AR predictor forecasts the spatio-temporal evolution of driving scene. We also equip the causal transformer with 3D RoPE across (t, y, x) coordinates for improved spatial and temporal reasoning.

An action injector fuses the action embedding $\mathbf{e}_a$ into the queries $\mathbf{q}^{(l)}$ at each layer l for action-conditioned fine-tuning:

$$\hat{\mathbf{q}}^{(l)} = \mathbf{q}^{(l)} + \gamma \cdot \mathrm{MLP}([\mathrm{LN}(\mathbf{q}^{(l)}) \parallel \mathbf{e}_a]), \tag{6}$$

where $\parallel$ denotes concatenation and γ is a zero-initialized learnable scale.

These action-conditioned queries then attend to the flattened historical latents $\mathbf{Z}_{\leq t}$ via cross-attention. By injecting 3D RoPE $(\mathbf{P}_q, \mathbf{P}_k)$, the network explicitly grounds temporal dynamics in physical space:

$$\tilde{\mathbf{q}}^{(l)} = \hat{\mathbf{q}}^{(l)} + \mathrm{CrossAttn}(\mathrm{LN}(\hat{\mathbf{q}}^{(l)}), \mathrm{LN}(\mathbf{Z}_{\leq t}), \mathrm{LN}(\mathbf{Z}_{\leq t}), \mathbf{P}_q, \mathbf{P}_k). \tag{7}$$

Finally, the last layer outputs a dense feature map $\mathbf{C}_{t+1} \in \mathbb{R}^{D_{AR} \times H_s \times W_s}$. This serves as the structural condition for the diffusion model and is linearly projected to predict the coarse future latent $\hat{\mathbf{Z}}_{t+1} \in \mathbb{R}^{D \times H_s \times W_s}$:

$$\mathbf{C}_{t+1} = \mathrm{LN}(\mathbf{q}^{(L)}), \quad \hat{\mathbf{Z}}_{t+1} = \mathrm{Proj}_{out}(\mathbf{C}_{t+1}). \tag{8}$$

Spatial Diffusion Refinement. Since we operate in a low-dimensional latent space and the AR model has already captured global dependencies, a lightweight 3-layer DiT head suffices to refine the coarse AR output into a high-fidelity future latent. We construct a token-wise spatial condition by adding the broadcasted timestep embedding $\mathbf{e}_\tau$ to the projected AR features:

$$\mathbf{C}_{\text{spatial}} = \text{Proj}(\mathbf{C}_{t+1}) + \mathbf{e}_\tau. \tag{9}$$

We then apply an adaptive modulation to each token $\mathbf{x}$ in the DiT blocks:

$$[\boldsymbol{\gamma}, \boldsymbol{\beta}] = \text{MLP}(\mathbf{C}_{\text{spatial}}), \quad \text{mod}(\mathbf{x}) = \text{LN}(\mathbf{x}) \odot (1 + \boldsymbol{\gamma}) + \boldsymbol{\beta}. \tag{10}$$

This token-level modulation ensures the DiT faithfully preserves the structural layout predicted by the AR model.

For the generation process, we adopt Rectified Flow [36]. Let $\mathbf{x}_1 = \boldsymbol{\mu}_{t+1}$ be the deterministic target latent and $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be standard Gaussian noise. The lightweight DiT $\mathbf{v}_\theta$ is trained to predict the constant velocity field driving the linear transition $\mathbf{x}_\tau = (1-\tau)\mathbf{x}_0 + \tau\mathbf{x}_1$:

$$\mathbf{v}_\theta(\mathbf{x}_\tau, \tau, \mathbf{C}_{\text{spatial}}) = \mathbf{x}_1 - \mathbf{x}_0, \tag{11}$$

where $\tau \in [0, 1]$ controls the transition from noise to target.

3.5 Training Objectives

We decouple the training process into a two-stage paradigm: representation compression and future latent generation.

Stage 1: Representation Compression. We train the compressor to encode unified scene representation into a compact latent space using a standard VAE objective. The loss comprises a MSE reconstruction term and a KL divergence regularization term:

$$\mathcal{L}_{\text{comp}} = \|\hat{\mathbf{S}} - \mathbf{S}\|_2^2 + \beta \cdot \text{KL}(q(\mathbf{Z}|\mathbf{S}) \| p(\mathbf{Z})), \tag{12}$$

where β is a very small weighting factor balancing the Gaussian prior $p(\mathbf{Z})$.

Stage 2: Future Latent Generation. In this stage, we optimize the generative models to predict future latents. For the AR model, we apply an MSE loss between the predicted coarse latent $\hat{\mathbf{Z}}_{t+1}$ and the ground truth $\mathbf{Z}_{t+1}$:

$$\mathcal{L}_{\text{AR}} = \|\hat{\mathbf{Z}}_{t+1} - \mathbf{Z}_{t+1}\|_2^2. \tag{13}$$

For the Spatial DiT, we train it to match the target velocity field under the Rectified Flow [36] paradigm. The diffusion loss is the MSE between the predicted velocity $\mathbf{v}_\theta$ and the ground-truth constant velocity:

$$\mathcal{L}_{\text{Diff}} = \|\mathbf{v}_\theta(\mathbf{x}_\tau, \tau, \mathbf{C}_{\text{spatial}}) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2. \tag{14}$$

The total generation loss is $\mathcal{L}_{\text{gen}} = \lambda_{\text{AR}}\mathcal{L}_{\text{AR}} + \mathcal{L}_{\text{Diff}}$, λ_{AR} is 0.5.

Table 1: Quantitative comparison on future image generation over the nuScenes validation set. Notably, our framework does not rely on any pre-trained LLMs or image generation models, while still producing diverse outputs across multiple modalities, demonstrating strong generalization capability.

Method	Pre-trained Model-Free	Volumetric Condition-Free	Multi-view	FID ↓
WoVoGen [39]	✗	✗	✓	27.60
X-Scene [59]	✗	✗	✓	<u>11.29</u>
OccScene [25]	✗	✗	✓	11.87
UniScene [24]	✗	✗	✓	**6.45**
DriveDreamer [48]	✗	✓	✗	14.90
DrivingGPT [5]	✗	✓	✗	<u>12.78</u>
Terra [1]	✗	✓	✗	26.73
Vista [9]	✗	✓	✗	**6.90**
Doe-1 [69]	✗	✓	✗	15.90
BEVControl [58]	✗	✓	✓	24.85
MagicDrive [8]	✗	✓	✓	16.20
Panacea [53]	✗	✓	✓	16.96
Drive-WM [49]	✗	✓	✓	<u>15.80</u>
DriveDreamer-2 [65]	✗	✓	✓	**11.20**
DiST-4D [12]	✓	✗	✓	**7.40**
DriveGAN [23]	✓	✓	✗	73.40
DrivingWorld [18]	✓	✓	✗	**7.40**
Epona [62]	✓	✓	✗	<u>7.50</u>
BEVGen [44]	✓	✓	✓	25.54
GenAD [68]	✓	✓	✓	<u>15.40</u>
MagicDrive-V2 [7]	✓	✓	✓	20.91
Doe-2	✓	✓	✓	**10.48**

4 Experiments

4.1 Experimental Setup

Dataset and metrics. All experiments are conducted on nuScenes [3] and Occ3D-nuScenes [45] benchmarks. The nuScenes dataset contains 1000 driving scenes with rich annotations, including multi-view images, 3D bounding boxes, and HD maps. To evaluate 3D scene understanding, we utilize the Occ3D-nuScenes dataset, which offers 3D occupancy labels for 18 categories based on the nuScenes dataset. The annotations cover a spatial range of $[-40m, -40m, -1m, 40m, 40m, 5.4m]$ with a voxel resolution of $0.4m$, resulting in a $200 \times 200 \times 16$ occupancy grid per frame. Following the standard protocol, both datasets are split into 700 training, 150 validation, and 150 test driving sequences.

For the temporal image generation task, we adopt the Fréchet Inception Distance (FID) as the evaluation metric. In 4D occupancy forecasting task, we report both the geometric Intersection over Union (IoU) and the semantic mean

Table 2: Quantitative results for depth forecasting. Our Doe-2 significantly outperforms all baselines across all metrics, achieving the lowest absolute relative error and the highest threshold accuracies.

Method	Abs. Rel. ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑
SurroundDepth [51]	0.28	0.66	0.84
M^2Depth [74]	<u>0.26</u>	<u>0.73</u>	<u>0.87</u>
Dist-4D [12]	0.39	0.58	0.81
OmniNWM [26]	0.23	0.81	0.93
Doe-2	**0.15**	**0.87**	**0.94**

Intersection over Union (mIoU). Additionally, for the motion planning evaluation, we employ the L2 displacement error and the collision rate for the predicted trajectories, which are sampled at 2Hz over a 3-second horizon. Finally, in the depth forecasting task, we evaluate the performance using the absolute relative error (AbsRel) and the percentage of pixels satisfying the threshold $\delta < 1.25$.

Implementation Details. We utilize pretrained DriveTok [73] as the scene encoder, processing 6 multi-view images with a resolution of 256×704. Specifically, we employ a DINOv3-ViTB [43] to extract multi-level features, which are lifted into a dense 3D representation of shape $32 \times 32 \times 768$. The compressor then encodes this into a compact latent space of $D = 64$. Our action-conditioned generative module comprises a 24-layer AR predictor with a hidden dimension of 384 and a lightweight 3-layer DiT with a hidden dimension of 3072. During the training phase, we optimize our model using the AdamW [37] optimizer with a weight decay of 0.01 and a batch size of 1 per GPU across 8 NVIDIA A800 GPUs, following a progressive two-stage paradigm. First, the compressor is trained for 5 epochs with a learning rate of 1×10^{-4}, involving 45M trainable parameters. Second, the generative module is trained for 100 epochs with a learning rate of 5×10^{-5} and 590M trainable parameters.

4.2 Temporal Image Generation

We evaluate the temporal visual synthesis performance of Doe-2 on the nuScenes [3] validation set. As shown in Fig. 5, Doe-2 generates temporally coherent and geometrically consistent multi-view future images. Quantitative results in Tab. 1 further show that Doe-2 achieves competitive image generation quality with an FID of 10.48. More importantly, unlike many prior approaches that rely on heavy volumetric conditions such as semantic occupancy or point clouds, or use pretrained LLMs and image generation models, Doe-2 is trained from scratch on a relatively small dataset containing only 700 scenes while remaining both volumetric condition-free and multi-view capable. This demonstrates the effectiveness and generalization ability of the proposed unified 3D scene representation.

In addition to image quality, we also evaluate the geometric fidelity of the generated multi-view depth using LiDAR-projected ground truth. As reported in Tab. 2, Doe-2 significantly outperforms existing baselines across all depth forecasting metrics, reducing Abs. Rel. error to 0.15 and improving $\delta < 1.25$ and

Table 3: Quantitative results for 4D occupancy forecasting on Occ3D-nuScenes. Avg. denotes the average performance of that in 1s, 2s, and 3s. OccWorld-D uses predictions from TPVFormer [20]. "-STC" and "-F" indicate models incorporating occupancy results from STCOcc [34] and FB-Occ [31], respectively.

Method	Input	Semantic mIoU (%) ↑				Geometric IoU (%) ↑				FPS ↑
		1s	2s	3s	Avg.	1s	2s	3s	Avg.	
OccWorld-O [67]	Occ GT	25.78	15.14	10.51	17.14	34.63	25.07	20.18	26.63	18.00
OccLLaMA-O [50]	Occ GT	25.05	19.49	15.26	19.93	34.56	28.53	24.41	29.17	-
RenderWorld [57]	Occ GT	28.69	18.89	14.83	20.89	37.74	28.41	24.08	30.08	-
Occ-LLM [55]	Occ GT	24.02	21.65	17.29	20.99	36.65	32.14	28.77	35.52	-
DFIT-OccWorld [61]	Occ GT	31.68	21.29	15.19	22.71	40.28	31.24	25.29	32.27	-
DOME [11]	Occ GT	35.11	25.89	20.29	27.10	43.99	35.36	29.74	36.36	6.54
UniScene [24]	Occ GT	35.37	29.59	25.08	31.76	38.34	32.70	29.09	34.84	-
I^2-World [35]	Occ GT	**47.62**	**38.58**	**32.98**	**39.73**	**54.29**	**49.43**	**45.69**	**49.80**	**37.04**
RenderWorld [57]	Camera	2.83	2.55	2.37	2.58	14.61	13.61	12.98	13.73	-
OccWorld-D [67]	Camera	11.55	8.10	6.22	8.62	18.90	16.26	14.43	16.53	-
OccLLaMA [50]	Camera	10.34	8.66	6.98	8.66	25.81	23.19	19.97	22.99	-
PreWorld [28]	Camera	12.27	9.24	7.15	9.55	23.62	21.62	19.63	21.62	-
Occ-LLM [55]	Camera	11.28	10.21	9.13	10.21	27.11	24.07	20.19	23.79	-
DFIT-OccWorld [61]	Camera	13.38	10.16	7.96	10.50	19.18	16.85	25.02	17.02	-
SparseWorld [6]	Camera	14.93	13.15	11.51	13.20	22.96	22.10	21.05	22.03	**8.0**
DOME-STC [11]	Camera	17.79	14.23	11.58	14.53	26.39	23.20	20.42	23.33	2.75
I^2-World-STC [35]	Camera	21.67	18.78	16.47	18.97	30.55	28.76	26.99	28.77	4.21
DOME-F [11]	Camera	24.12	17.41	13.24	18.25	35.18	27.90	23.44	28.84	-
DTT [54]	Camera	24.87	18.30	15.63	19.60	38.98	37.45	31.89	36.11	-
COME-F [42]	Camera	26.56	21.73	18.49	22.26	48.08	43.84	40.28	44.07	-
Doe-2	Camera	**30.90**	**25.70**	**22.44**	**26.29**	**64.02**	**60.59**	**57.75**	**60.73**	~ 4.0

$\delta < 1.25^2$ to 0.87 and 0.94, respectively. These results indicate that performing generation in an explicit unified 3D representation not only improves photorealistic temporal image synthesis but also preserves accurate metric-scale scene geometry, effectively bridging visual generation and 3D spatial reasoning.

4.3 4D Occupancy Forecasting

To validate the geometric awareness of Doe-2, we assess its performance on 4D occupancy forecasting. Tab. 3 and Fig. 5 show that Doe-2 achieves impressive geometric forecasting accuracy. Crucially, Doe-2 achieves accurate 3D geometric prediction alongside high-fidelity visual synthesis in a single pass, thereby alleviating the cascaded errors and computational overhead in multi-stage pipelines.

4.4 Controllable Generation

Agile control responsiveness is a hallmark of an effective autonomous driving world model. We investigate the ability of Doe-2 to simulate diverse, action-conditioned future rollouts. By injecting different ego-actions (e.g., accelerating, braking, turning left/right) into the same historical context, we observe the evolution of the predicted scene as shown in Fig. 6. The visualizations show that Doe-2 can generate highly controllable future scenes. These results indicate that our action-conditioned AR architecture effectively models the structural layout

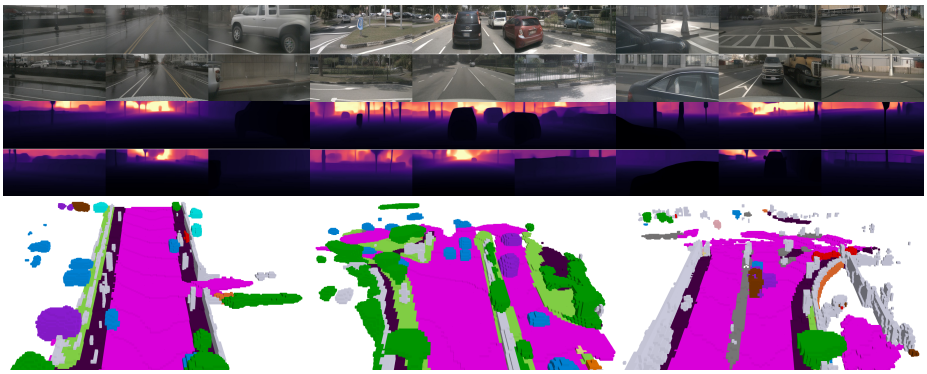

Fig. 5: Multimodal generation results of Doe-2. By decoding from a unified and compact 3D latent representation, our method is able to simultaneously generate high-fidelity multi-view RGB images, depth maps, semantic segmentation, and 3D occupancy in a single forward pass.

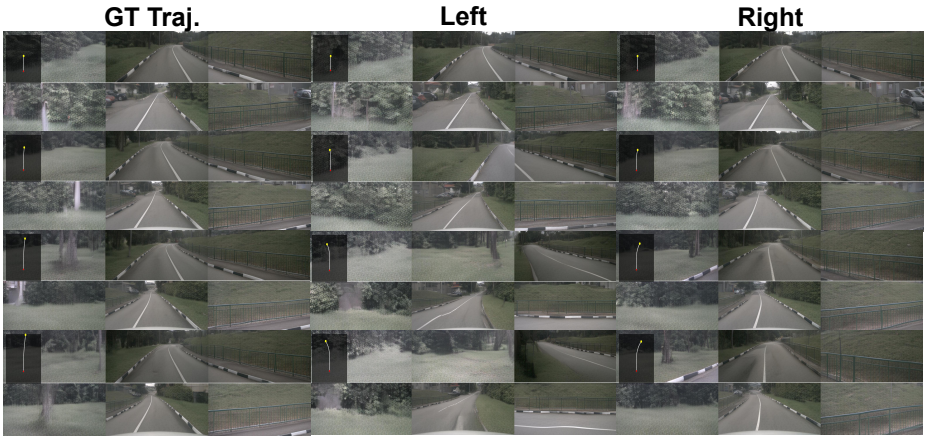

Fig. 6: Visualization of controllable generation. The model predicts different future scenes based on specific action instructions: following the GT trajectory, turning left, and turning right.

of future latents under explicit control signals, demonstrating the potential of Doe-2 to serve as a simulator for driving policy evaluation.

4.5 Motion Planning

We further evaluate the motion planning capability of Doe-2 on the nuScenes [3] dataset. To generate safe and temporally consistent ego-actions, we adopt a diffusion-based motion head following DiffusionDrive [33], which predicts future ego-trajectories directly from the generated unified 3D representations.

As shown in Tab. 4, Doe-2 significantly reduces the L2 trajectory error compared with both end-to-end planning baselines and previous world-model-based methods. It is worth noting that many existing end-to-end planners rely on high-level auxiliary supervision, such as *HD Maps* and *Agent Motion*, which provide explicit cues for lane structure, agent interaction, and collision avoidance. In

Table 4: Motion planning performance. Aux. Sup. denotes auxiliary supervision apart from the ego trajectory. "-F" indicate models incorporating occupancy results from FB-Occ [31]. We use bold and underlined numbers to denote the best and second-best results, respectively. † denotes using the metric computation adopted in VAD [21].

Method	Input	Aux. Sup.	L2 (m) ↓				Collision Rate (%) ↓				FPS
			1s	2s	3s	Avg.	1s	2s	3s	Avg.	
IL [41]	LiDAR	None	0.44	1.15	2.47	1.35	0.08	0.27	1.95	0.77	-
NMP [60]	LiDAR	Box & Motion	0.53	1.25	2.67	1.48	0.04	0.12	0.87	0.34	-
FF [16]	LiDAR	Freespace	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43	-
EO [22]	LiDAR	Freespace	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33	-
OccNet [46]	3D-Occ	Map & Box	1.29	2.31	2.98	2.25	0.20	0.56	1.30	0.69	-
OccWorld-O [67]	3D-Occ	None	0.43	1.08	1.99	1.17	0.07	0.38	1.35	0.60	**18.0**
RenderWorld-O [57]	3D-Occ	None	0.35	0.91	1.84	1.03	0.05	0.40	1.39	0.61	-
OccLLaMA-O [50]	3D-Occ	None	0.35	0.91	1.84	1.03	0.05	0.40	1.39	0.61	-
DFIT-OccWorld [61]	3D-Occ	None	0.38	0.96	1.73	1.02	0.07	0.39	0.90	0.45	-
ST-P3 [17]	Camera	Map & Box & Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	1.6
UniAD [19]	Camera	Map & Box & Motion & Tracklets & Occ	0.48	0.96	1.65	1.03	0.05	0.17	**0.71**	**0.31**	1.8
VAD-Tiny [21]	Camera	Map & Box & Motion	0.60	1.23	2.06	1.30	0.31	0.53	1.33	0.72	16.8
VAD-Base [21]	Camera	Map & Box & Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53	4.5
GenAD [68]	Camera	Map & Box & Motion	0.36	0.83	1.55	0.91	0.06	0.23	1.00	0.43	-
OccNet [46]	Camera	3D-Occ & Map & Box	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72	2.6
OccWorld-D [67]	Camera	3D-Occ	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87	2.8
RenderWorld [57]	Camera	3D-Occ	0.48	1.30	2.67	1.48	0.14	0.55	2.23	0.97	-
OccLLaMA-F [50]	Camera	3D-Occ	0.38	1.07	2.15	1.20	0.06	0.39	1.65	0.70	-
DFIT-OccWorld [61]	Camera	3D-Occ	0.42	1.14	2.19	1.25	0.09	0.19	1.37	0.55	-
OmniDrive [47]	Camera	Map & Box & Language	0.40	0.80	**1.32**	0.84	0.04	0.46	2.32	0.94	-
Doe-1 [69]	Camera	Language	0.50	1.18	2.11	1.26	0.04	0.37	1.19	0.53	-
UniUGP [38]	Camera	Language	0.58	1.14	1.95	1.23	0.01	0.19	0.81	0.33	-
Doe-2	Camera	3D-Occ & Depth	**0.23**	**0.65**	1.46	**0.78**	**0.00**	**0.08**	1.09	0.39	∼ 6.0
VAD-Tiny† [21]	Camera	Map & Box & Motion	0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38	16.8
VAD-Base† [21]	Camera	Map & Box & Motion	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22	4.5
OccWorld-D† [67]	Camera	3D-Occ	0.39	0.73	1.18	0.77	0.11	0.19	0.67	0.32	2.8
GenAD† [68]	Camera	Map & Box & Motion	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19	-
Doe-2†	Camera	3D-Occ & Depth	**0.17**	**0.35**	**0.64**	**0.39**	**0.00**	**0.05**	**0.32**	**0.12**	∼ **6.0**

contrast, Doe-2 relies only on geometric supervision, without using high-level planning priors. Despite this more challenging setting, Doe-2 still achieves competitive planning performance. These results demonstrate that the unified 3D world model effectively captures scene evolution, spatial geometry, and semantic relationships, allowing the learned representations to directly support more accurate, safe, and robust trajectory planning in complex driving scenarios.

4.6 Ablation Study

We conduct comprehensive ablation studies on the nuScenes [3] dataset to validate our core architectural designs and hyperparameters of Doe-2.

Impact of Compression Ratios. To alleviate the computational overhead of 3D scene generation, we compress the high-dimensional tokens into a compact latent space. We investigate the trade-off between generation quality and inference latency by varying the channel compression ratios. As shown in Tab. 5, operating directly on uncompressed tokens not only leads to prohibitive inference latency but also inherently degrades the synthesis quality of the DiT model due to the high-dimensional representation spaces. Conversely, an excessively

Table 5: Performance of Doe-2 under different compression rates on SurroundOcc [52]. While higher latent dimensions (e.g., 768) slightly improve reconstruction fidelity, they severely hinder generation quality due to the increased difficulty of modeling high-dimensional distributions. A latent dimension of 64 strikes the optimal balance, achieving the best generation performance and high efficiency (approximately 4.0 FPS) with minimal loss in reconstruction.

Method	Latent Dim	Reconstruction		Generation		FPS
		mIoU	IoU	mIoU	IoU	
Doe-2	768	**20.02**	**33.08**	8.24	23.07	~ 2.3
Doe-2	128	19.88	32.21	10.29	27.84	~ 3.4
Doe-2	64	19.65	31.84	**12.04**	**30.56**	~ 4.0
Doe-2	32	19.01	30.27	9.33	25.12	~ 4.2

Table 6: Performance of Doe-2 with different β settings on SurroundOcc [52] dataset. Setting $\beta = 0.00001$ achieves the optimal balance, providing a well-regularized latent space that results in the highest generation mIoU and IoU.

Method	KL weight β	Reconstruction		Generation	
		mIoU	IoU	mIoU	IoU
Doe-2	0.01	18.29	29.77	8.92	24.84
Doe-2	0.00001	19.65	31.84	**12.04**	**30.56**
Doe-2	0.000001	**19.73**	**32.41**	11.45	30.17

high compression ratio results in the loss of fine-grained texture and geometric details, leading to degraded 3D occupancy metrics. Our chosen compression ratio strikes an optimal balance, drastically accelerating the inference speed (FPS) while preserving holistic 3D scene perception and high-fidelity generation.

Regularization with KL Weight ($\boldsymbol{\beta}$). The KL divergence penalty weight (β) in our feature compressor plays a crucial role in shaping the latent space for downstream generative modeling. We evaluate the generation results under different β values. As shown in Tab. 6, When β is set too high, the model suffers from severe posterior collapse, yielding overly smoothed latents that lack meaningful spatial-temporal and semantic information. On the other hand, when β is entirely removed or set too low, the latent space becomes highly irregular. While reconstruction remains sharp, the downstream AR and DiT models struggle to fit this unconstrained distribution, resulting in severe degradation during generation. An appropriately small β effectively regularizes the latent space into a smooth, continuous manifold highly amenable to generative models.

Hybrid Generative Architecture. To justify our proposed action-conditioned hybrid paradigm, we compare Doe-2 against pure AR and pure DiT baselines. As shown in Tab. 7 and Fig. 7, while the pure AR model successfully captures global spatio-temporal dynamics, its limited capacity for high-frequency detail synthesis yields blurry results. Conversely, a full-scale DiT incurs prohibitive computational overhead. Even after reducing its hidden dimension width from 3072 to 1536 to maintain network depth, the parameter count remains over twice

Table 7: Performance of Doe-2 with different generation architectures on the nuScenes [3] dataset. Our AR+DiT architecture combines the strengths of both AR and diffusion-based methods, enabling high-fidelity and controllable generation.

Method	Gen. Arch.	Generation			FPS ↑
		mIoU ↑	IoU ↑	gFID ↓	
Doe-2	Pure AR	11.79	30.01	29.93	∼6.0
Doe-2	Pure DiT	10.23	29.27	16.25	∼1.9
Doe-2	AR+DiT Head	**12.04**	**30.56**	**10.48**	∼4.0

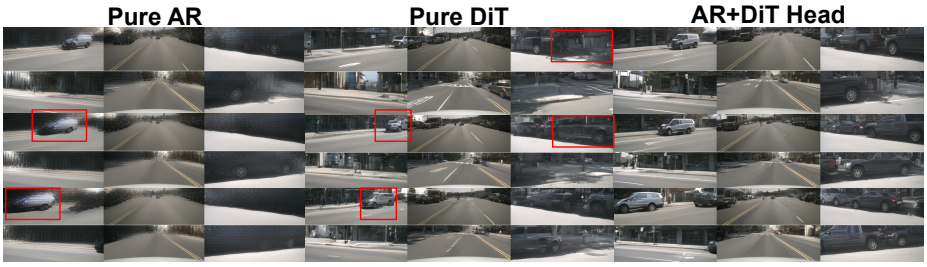

Fig. 7: Ablation study for generation architecture. The pure AR model captures dynamic evolution well but suffers from blurry visuals. The pure DiT model offers high image quality but struggles with dynamic object modeling, resulting in artifacts. Our AR+DiT architecture combines the strengths of both, enabling high-fidelity and controllable dynamic generation. **Red boxes** indicate generation errors.

that of our hybrid architecture. Furthermore, training such a massive DiT from scratch on limited datasets like nuScenes [3] leads to data starvation, degrading overall generation quality and failing to model dynamic objects. Our Doe-2 resolves these issues: by utilizing the AR model for reliable global spatio-temporal forecasting and appending a lightweight DiT head solely for spatial refinement, it efficiently delivers high-fidelity, dynamic scene synthesis.

5 Conclusion

In this work, we presented Doe-2, a native 3D representation world model tailored for unified driving scene forecasting. By compressing multi-view inputs into a highly compact 3D latent space, Doe-2 effectively bridges the gap between holistic geometric awareness and efficient generative modeling. Our action-conditioned architecture, which pairs an autoregressive model for global spatio-temporal dynamics with a lightweight DiT head for spatial refinement, enables the simultaneous generation of multi-view images, 3D occupancy, depth, and semantics in a single pass. This unified paradigm significantly mitigates compounding errors and boosts inference efficiency, demonstrating immense potential for scalable closed-loop simulation.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62125603, Grant 62576188, Grant 62336004, and Grant 62321005, and in part by the Beijing Natural Science Foundation under Grant No. L247009.

References

1. Arai, H., Ishihara, K., Takahashi, T., Yamaguchi, Y.: Act-bench: Towards action controllable world models for autonomous driving. arXiv preprint arXiv:2412.05337 (2024)
2. Baldassarre, F., Szafraniec, M., Terver, B., Khalidov, V., Massa, F., LeCun, Y., Labatut, P., Seitzer, M., Bojanowski, P.: Back to the features: Dino as a foundation for video world models. arXiv preprint arXiv:2507.19468 (2025)
3. 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)
4. Chen, A., Zheng, W., Wang, Y., Zhang, X., Zhan, K., Jia, P., Keutzer, K., Zhang, S.: Geodrive: 3d geometry-informed driving world model with precise action control. arXiv preprint arXiv:2505.22421 (2025)
5. Chen, Y., Wang, Y., Zhang, Z.: Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. arXiv preprint arXiv:2412.18607 (2024)
6. Dang, C., Liu, H., Bao, J., An, P., Tang, X., Ma, J., Sun, B., Wang, Y., et al.: Sparseworld: A flexible, adaptive, and efficient 4d occupancy world model powered by sparse and dynamic queries. arXiv preprint arXiv:2510.17482 (2025)
7. Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. arXiv preprint arXiv:2411.13807 (2024)
8. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601 (2023)
9. 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)
10. Gao, Y., Chen, C., Chen, T., Gu, J.: One layer is enough: Adapting pretrained visual encoders for image generation. arXiv preprint arXiv:2512.07829 (2025)
11. Gu, S., Yin, W., Jin, B., Guo, X., Wang, J., Li, H., Zhang, Q., Long, X.: Dome: Taming diffusion model into high-fidelity controllable occupancy world model. arXiv preprint arXiv:2410.10429 (2024)
12. Guo, J., Ding, Y., Chen, X., Chen, S., Li, B., Zou, Y., Lyu, X., Tan, F., Qi, X., Li, Z., et al.: Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27231–27241 (2025)
13. Guo, X., Wu, Z., Xiong, K., Xu, Z., Zhou, L., Xu, G., Xu, S., Sun, H., Wang, B., Chen, G., et al.: Genesis: Multimodal driving scene generation with spatio-temporal and cross-modal consistency. arXiv preprint arXiv:2506.07497 (2025)
14. Hassan, M., Stapf, S., Rahimi, A., Rezende, P., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., et al.: Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22404–22415 (2025)
15. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)

16. Hu, P., Huang, A., Dolan, J., Held, D., Ramanan, D.: Safe local motion planning with self-supervised freespace forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12732–12741 (2021)
17. 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)
18. Hu, X., Yin, W., Jia, M., Deng, J., Guo, X., Zhang, Q., Long, X., Tan, P.: Drivingworld: Constructing world model for autonomous driving via video gpt. arXiv preprint arXiv:2412.19505 (2024)
19. 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)
20. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9223–9232 (2023)
21. 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)
22. Khurana, T., Hu, P., Dave, A., Ziglar, J., Held, D., Ramanan, D.: Differentiable raycasting for self-supervised occupancy forecasting. In: European Conference on Computer Vision. pp. 353–369. Springer (2022)
23. Kim, S.W., Philion, J., Torralba, A., Fidler, S.: Drivegan: Towards a controllable high-quality neural simulation. In: CVPR (2021)
24. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. In: Proceedings of the computer vision and pattern recognition conference. pp. 11971–11981 (2025)
25. Li, B., Jin, X., Wang, J., Shi, Y., Sun, Y., Wang, X., Ma, Z., Xie, B., Ma, C., Yang, X., et al.: Occscene: Semantic occupancy-based cross-task mutual learning for 3d scene generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
26. Li, B., Ma, Z., Du, D., Peng, B., Liang, Z., Liu, Z., Ma, C., Jin, Y., Zhao, H., Zeng, W., et al.: Omninwm: Omniscient driving navigation world models. arXiv preprint arXiv:2510.18313 (2025)
27. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
28. Li, X., Li, P., Zheng, Y., Sun, W., Wang, Y., Chen, Y.: Semi-supervised vision-centric 3d occupancy world model for autonomous driving. arXiv preprint arXiv:2502.07309 (2025)
29. 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)
30. 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)
31. Li, Z., Yu, Z., Austin, D., Fang, M., Lan, S., Kautz, J., Alvarez, J.M.: Fb-occ: 3d occupancy prediction based on forward-backward view transformation. arXiv preprint arXiv:2307.01492 (2023)

32. Liang, D., Zhang, D., Zhou, X., Tu, S., Feng, T., Li, X., Zhang, Y., Du, M., Tan, X., Bai, X.: Unifuture: A 4d driving world model for future generation and perception. In: IEEE International Conference on Robotics and Automation. IEEE (2026)
33. 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)
34. Liao, Z., Wei, P., Chen, S., Wang, H., Ren, Z.: Stcocc: Sparse spatial-temporal cascade renovation for 3d occupancy and scene flow prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1516–1526 (2025)
35. Liao, Z., Wei, P., Zhang, R., Chen, S., Wang, H., Ren, Z.: I2-world: Intra-inter tokenization for efficient dynamic 4d scene forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25810–25819 (2025)
36. 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)
37. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
38. Lu, H., Liu, Z., Jiang, G., Luo, Y., Chen, S., Zhang, Y., Chen, Y.C.: Uniugp: Unifying understanding, generation, and planing for end-to-end autonomous driving. arXiv preprint arXiv:2512.09864 (2025)
39. Lu, J., Huang, Z., Zhang, J., Yang, Z., Zhang, L.: Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. arXiv preprint arXiv:2312.02934 (2023)
40. Min, C., Zhao, D., Xiao, L., Zhao, J., Xu, X., Zhu, Z., Jin, L., Li, J., Guo, Y., Xing, J., et al.: Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15522–15533 (2024)
41. Ratliff, N.D., Bagnell, J.A., Zinkevich, M.A.: Maximum margin planning. In: Proceedings of the 23rd international conference on Machine learning. pp. 729–736 (2006)
42. Shi, Y., Jiang, K., Meng, Q., Wang, K., Wang, J., Sun, W., Wen, T., Yang, M., Yang, D.: Come: Adding scene-centric forecasting control to occupancy world model. arXiv preprint arXiv:2506.13260 (2025)
43. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
44. Swerdlow, A., Xu, R., Zhou, B.: Street-view image generation from a bird's-eye view layout. IEEE Robotics and Automation Letters (2024)
45. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. Advances in Neural Information Processing Systems **36**, 64318–64330 (2023)
46. 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)
47. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: Proceedings of the computer vision and pattern recognition conference. pp. 22442–22452 (2025)
48. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)

49. 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)
50. Wei, J., Yuan, S., Li, P., Hu, Q., Gan, Z., Ding, W.: Occllama: An occupancy-language-action generative world model for autonomous driving. arXiv preprint arXiv:2409.03272 (2024)
51. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Rao, Y., Huang, G., Lu, J., Zhou, J.: Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In: Conference on robot learning. pp. 539–549. PMLR (2023)
52. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21729–21740 (2023)
53. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: CVPR (2024)
54. Xu, H., Peng, P., Tan, G., Chang, Y., Zhao, Y., Tian, Y.: Delta-triplane transformers as occupancy world models. arXiv preprint arXiv:2503.07338 (2025)
55. Xu, T., Lu, H., Yan, X., Cai, Y., Liu, B., Chen, Y.: Occ-llm: Enhancing autonomous driving with occupancy-based large language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8434–8441. IEEE (2025)
56. Yan, T., Wu, D., Han, W., Jiang, J., Zhou, X., Zhan, K., Xu, C.z., Shen, J.: Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27531–27541 (2025)
57. Yan, Z., Dong, W., Shao, Y., Lu, Y., Liu, H., Liu, J., Wang, H., Wang, Z., Wang, Y., Remondino, F., et al.: Renderworld: World model with self-supervised 3d label. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6063–6070. IEEE (2025)
58. Yang, K., Ma, E., Peng, J., Guo, Q., Lin, D., Yu, K.: Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. arXiv preprint arXiv:2308.01661 (2023)
59. Yang, Y., Liang, A., Mei, J., Ma, Y., Liu, Y., Lee, G.H.: X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. arXiv preprint arXiv:2506.13558 (2025)
60. Zeng, W., Luo, W., Suo, S., Sadat, A., Yang, B., Casas, S., Urtasun, R.: End-to-end interpretable neural motion planner. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8660–8669 (2019)
61. Zhang, H., Xue, Y., Yan, X., Zhang, J., Qiu, W., Bai, D., Liu, B., Cui, S., Li, Z.: An efficient occupancy world model via decoupled dynamic flow and image-assisted training. arXiv preprint arXiv:2412.13772 (2024)
62. 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)
63. Zhang, Y., Gong, S., Xiong, K., Ye, X., Tan, X., Wang, F., Huang, J., Wu, H., Wang, H.: Bevworld: A multimodal world model for autonomous driving via unified bev latent space (2024)
64. Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., Huang, G., Chen, X., Wang, B., Zhang, Y., 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 (2025)

65. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. arXiv preprint arXiv:2403.06845 (2024)
66. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
67. 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)
68. 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)
69. 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)
70. Zhou, H., Lin, L., Wang, J., Lu, Y., Bai, D., Liu, B., Wang, Y., Geiger, A., Liao, Y.: Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. IEEE TPAMI (2025)
71. Zhou, X., Liang, D., Chen, X., Tan, F., Zhang, D., Zhao, H., Bai, X.: Hermes++: Toward a unified driving world model for 3d scene understanding and generation. arXiv preprint arXiv:2604.28196 (2026)
72. Zhou, X., Liang, D., Tu, S., Chen, X., Ding, Y., Zhang, D., Tan, F., Zhao, H., Bai, X.: Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27817–27827 (2025)
73. Zhuo, D., Zheng, W., Zuo, S., Yan, S., Hou, L., Zhou, J., Lu, J.: Drivetok: 3d driving scene tokenization for unified multi-view reconstruction and understanding. arXiv preprint arXiv:2603.19219 (2026)
74. Zou, Y., Ding, Y., Qiu, X., Wang, H., Zhang, H.: M 2 depth: Self-supervised two-frame m ulti-camera m etric depth estimation. In: European Conference on Computer Vision. pp. 269–285. Springer (2024)