UniDriveDreamer: A Single-Stage Multimodal World Model for Autonomous Driving

Guosheng Zhao[1,2*], Yaozeng Wang[1*], Xiaofeng Wang[1*], Zheng Zhu[1*✉], Yu Tingdong[1], Guan Huang[1], Yongchen Zai[3], Ji Jiao[3], Changliang Xue[3], Xiaole Wang[3], Zhen Yang[3], Futang Zhu[3], and Xingang Wang[2✉]

[1] GigaAI, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] BYD, Guangdong, China

Abstract. World models have demonstrated significant promise for data synthesis in autonomous driving. However, existing methods predominantly concentrate on single-modality generation, typically focusing on either multi-camera video or LiDAR sequence synthesis. In this paper, we propose *UniDriveDreamer*, a single-stage unified multimodal world model for autonomous driving, which directly generates multimodal future observations without relying on intermediate representations or cascaded modules. Our framework introduces a LiDAR-specific variational autoencoder (VAE) designed to encode input LiDAR sequences, alongside a video VAE for multi-camera images. To ensure cross-modal compatibility and training stability, we propose Unified Latent Anchoring (ULA), which explicitly aligns the latent distributions of the two modalities. The aligned features are fused and processed by a diffusion transformer that jointly models their geometric correspondence and temporal evolution. Additionally, structured scene layout information is projected per modality as a conditioning signal to guide the synthesis. Extensive experiments demonstrate that *UniDriveDreamer* outperforms previous state-of-the-art methods in both video and LiDAR generation, while also yielding measurable improvements in downstream driving tasks.

Keywords: Autonomous Driving · Multimodal Generation

1 Introduction

World models [1, 3, 12, 15, 19, 27–29, 34, 38, 41, 43, 50, 51] have made significant progress in autonomous driving, and their powerful data synthesis capabilities offer a promising solution for reducing the costs of data acquisition and annotation. Recent studies [10, 11, 31, 42, 46] have established a notable paradigm that leverages structured scene layouts, such as 3D bounding boxes and HDMaps,

* These authors contributed equally to this work.
✉ Corresponding authors. zhengzhu@ieee.org, xingang.wang@ia.ac.cn.

as conditional inputs for driving data generation. The synthesized data can be further utilized to enhance the performance of downstream tasks [22, 25].

Nevertheless, existing methods primarily focus on single-modality generation, such as camera-only RGB synthesis [10–12, 21, 38, 41, 42, 45, 52] or LiDAR sequence generation [6, 16, 54, 55]. Although these approaches are able to produce highly realistic data that benefit downstream perception tasks, the inherent single-modality nature limits their applicability in scenarios demanding diverse multi-sensor observations. Recently, a few efforts have explored multimodal generation for autonomous driving, demonstrating the concrete potential of multimodal data synthesis to improve downstream task performance. However, existing methods [13, 20] typically decouple the generation of different modalities. In such frameworks, the synthesis of one modality (e.g., LiDAR) is explicitly conditioned on the output or latent representation of another (e.g., RGB). This decoupled architecture lacks explicit, bidirectional interaction between the RGB and LiDAR modalities during the synthesis process, which consequently undermines cross-modal consistency. The resulting unidirectional information flow impedes the effective utilization of complementary sensor data, consequently degrading the overall quality and coherence of the generated multimodal outputs.

To address this challenge, we introduce *UniDriveDreamer*, a single-stage multimodal world model designed for autonomous driving, which enables deep fusion and bidirectional interaction across different modalities, thereby substantially enhancing the spatiotemporal and cross-modal consistency of synthesized outputs. Specifically, a LiDAR-specific variational autoencoder (VAE) is designed to encode input range images into a latent space, while multi-view images are processed by a video VAE [17, 36]. Furthermore, to stabilize training and ensure compatibility between modalities, we propose Unified Latent Anchoring (ULA), which aligns the distribution of LiDAR latents with the prior of a pretrained RGB VAE via affine transformations derived from empirical statistics. The aligned latent features are then fused and processed by a diffusion transformer, which jointly models both intra- and cross-modal spatiotemporal relationships. In addition, we incorporate structured scene layout information to guide the generation process by projecting it into each modality as a conditioning signal. Extensive experimental results demonstrate that *UniDriveDreamer* surpasses previous state-of-the-art methods (SOTA) in multimodal data synthesis quality. Specifically, for video generation, *UniDriveDreamer* achieves an FID of 2.81 and an FVD of 11.44. For LiDAR synthesis, it attains an MMD of 0.27 and a JSD of 0.039, corresponding to 82.3% and 45.8% relative improvements over UniScene [20]. In addition, downstream task evaluations further confirm the practical effectiveness of the generated data, showing relative improvements of +1.2% in mAP and +0.7% in NDS.

Overall, our contributions can be summarized as follows:

- We propose *UniDriveDreamer*, a single-stage multimodal world model for autonomous driving, which directly generates temporally consistent and geometrically coherent future observations, including multi-camera videos and

LiDAR sweeps, without relying on intermediate representations or cascaded modules.
- We propose a LiDAR-specific VAE, capable of accurately reconstructing LiDAR sequences. Furthermore, to ensure cross-modal compatibility, Unified Latent Anchoring (ULA) is proposed to align their distributions via affine transformations based on empirical statistics, thereby stabilizing training and ensuring cross-modal compatibility.
- Experimental results demonstrate that *UniDriveDreamer* not only surpasses previous SOTA methods in multimodal synthesis quality but also delivers significant improvements in downstream perception task performance.

2 Related Work

2.1 Unimodal Driving Scene Generation

The rapid advancement of diffusion models [1, 3–5, 18, 30, 36, 37, 39, 47, 53] has catalyzed a surge of research in driving video generation [2, 10–12, 15, 21, 29, 31, 32, 38, 42, 46, 52]. These methods commonly employ structured scene layouts as conditional inputs to synthesize realistic driving videos, substantially lowering the cost of data acquisition and annotation while demonstrating effectiveness in enhancing downstream perception performance. Furthermore, several approaches [26, 41, 44] have investigated the application of world models in end-to-end autonomous driving. In parallel, as LiDAR serves as a crucial sensing modality in driving scenarios, recent methods [6, 16, 29, 54, 55] have begun to introduce generative techniques for LiDAR data synthesis. These methods typically represent LiDAR data as range images by transforming the 3D point clouds to 2D pixel space, thereby aligning its format with that of RGB images. This representation facilitates the application of well-established image and video generation techniques, leading to significant advancements in the field. Despite these advances, existing approaches remain predominantly focus on single-modality generation, thereby limiting the realization of sensor-complete autonomous driving simulation environments. To bridge this gap, we propose *UniDriveDreamer*, a unified world model designed to address multimodal data synthesis in driving scenarios, capable of jointly generating multi-view camera videos and LiDAR sequences.

2.2 Multimodal Driving Scene Generation

To meet the demand for multimodal data in downstream tasks, recent works [13, 20, 23, 33, 49] have begun to explore multimodal data generation for driving scenarios. For instance, UniScene [20] proposes a two-stage pipeline that first generates an occupancy representation, then synthesizes camera and LiDAR data conditioned on it. Following a similar cascaded philosophy, Genesis [13] employs a dual-branch architecture, where the explicit occupancy is replaced by a latent video representation extracted from the RGB branch, and LiDAR generation is

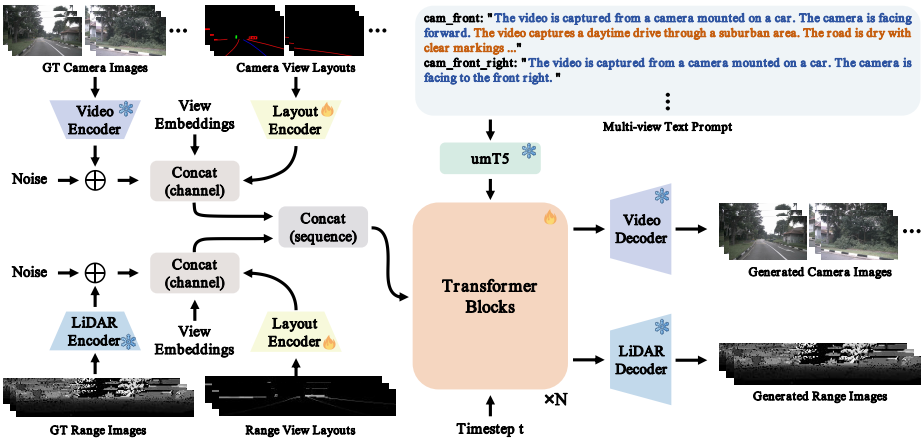

Fig. 1: The overall framework of *UniDriveDreamer*. Our *UniDriveDreamer* consists of four core components: (1) two modality-specific VAEs that encode multi-view camera images and LiDAR range maps into a shared latent space; (2) a layout encoder that projects structured scene layout information into corresponding latent representations; (3) a text encoder that encode multi-view text prompts into corresponding prompt embeddings; and (4) a diffusion transformer that jointly models spatiotemporal coherence within each modality and cross-modal consistency across modalities.

conditioned on these features sequentially. Despite these advances, such cascaded or sequential paradigms lack explicit bidirectional interaction between RGB and LiDAR during synthesis, which constrains the overall quality and consistency of the generated multimodal outputs. In contrast, our *UniDriveDreamer* adopts a single-stage architecture that enables bidirectional interaction between RGB and LiDAR modalities throughout the generation process, facilitating deeper fusion and more coherent multimodal synthesis.

3 UniDriveDreamer

The overall framework is illustrated in Fig. 1. Multi-view images and LiDAR range maps are first encoded by their dedicated VAEs. The resulting latent features are then concatenated in the channel dimension with the corresponding modality-specific layout latent features and view embeddings. Next, the two sets of multimodal latent tokens are concatenated along the sequence dimension and fed directly into a diffusion transformer, which jointly models both intra-modal spatiotemporal consistency and cross-modal alignment in the latent space. Finally, the output latent representations are decoded back into their respective modalities, RGB video frames and LiDAR range sequences, through the corresponding decoder networks.

3.1 LiDAR VAE

In this section, we detail the training and inference process of the proposed LiDAR-specific VAE. Notably, to facilitate the joint generation of LiDAR sequences and camera videos, we adopt range maps as the LiDAR representation, since their image-like format is analogous to RGB images and thus more suitable for unifying the two modalities. For LiDAR generation, there is currently no definitive paradigm in the community, while range-map-based representations have been widely adopted by existing methods, such as RangeLDM [16] and Cosmos-Drive-Dreams [29]. Inspired by these works, we choose range maps to better support joint multimodal generation within a unified framework.

Training. Our LiDAR VAE preserves the same architectural backbone as the RGB VAE proposed in [36], differing only in the number of input and output channels, which are set to 1 instead of 3 to accommodate the single-channel nature of range images. Following [29], we further mitigate the inherent sparsity of range images by repeating each row of the range map four times, yielding an input sequence $v^{\mathrm{L}} \in \mathbb{R}^{(1+T)\times H^{\mathrm{L}}\times W^{\mathrm{L}}\times 1}$. The encoder of the LiDAR VAE maps the sequence to a latent representation $z^{\mathrm{L}} = \xi(v^{\mathrm{L}}) \in \mathbb{R}^{(1+\frac{T}{4})\times \frac{H^{\mathrm{L}}}{8}\times \frac{W^{\mathrm{L}}}{8}\times C}$, which is subsequently decoded back into the reconstructed LiDAR video $\hat{v}^{\mathrm{L}} = \mathcal{G}(z^{\mathrm{L}}) \in \mathbb{R}^{(1+T)\times H^{\mathrm{L}}\times W^{\mathrm{L}}\times 1}$. The model is trained using the following composite loss function:

$$\mathcal{L}_{\mathbf{vae}} = \lambda_1 \|v^{\mathrm{L}} - \hat{v}^{\mathrm{L}}\|_1 + \lambda_2 D_{\mathrm{KL}}(q(z^{\mathrm{L}}|v^{\mathrm{L}})\|p(z^{\mathrm{L}})) + \lambda_3 \mathcal{L}_{\mathrm{LPIPS}}(v^{\mathrm{L}}, \hat{v}^{\mathrm{L}}), \tag{1}$$

where $D_{\mathrm{KL}}(\cdot)$ is the Kullback–Leibler divergence between the approximate posterior $q(z^{\mathrm{L}}|v^{\mathrm{L}})$ and the standard Gaussian prior $p(z^{\mathrm{L}}) \sim \mathcal{N}(z^{\mathrm{L}}; \mathbf{0}, \mathbf{I})$, and $\mathcal{L}_{\mathrm{LPIPS}}$ is the Learned Perceptual Image Patch Similarity (LPIPS) loss [48]. Since LPIPS is originally defined for RGB images, we replicate the single-channel range images across three channels when computing this perceptual loss term.

Inference. Once trained, the LiDAR VAE is integrated into the multimodal world model for joint training. To stabilize training and ensure cross-modal compatibility, we introduce Unified Latent Anchoring (ULA), which aligns the latent distribution of LiDAR features with that of the camera modality. Let $\mu^{\mathrm{C}}, \sigma^{\mathrm{C}}$ be the normalization parameters of the pretrained RGB VAE, $\mu_1^{\mathrm{C}}, \sigma_1^{\mathrm{C}}$ denote the statistics computed from the current driving dataset using the RGB VAE, and $\mu_1^{\mathrm{L}}, \sigma_1^{\mathrm{L}}$ refer the corresponding statistics from the LiDAR VAE. The calibrated normalization parameters $\mu^{\mathrm{L}}, \sigma^{\mathrm{L}}$ for the LiDAR latent features are then obtained as:

$$\begin{aligned}
& \frac{\frac{z^{\mathrm{L}}-\mu_1^{\mathrm{L}}}{\sigma_1^{\mathrm{L}}} \cdot \sigma_1^{\mathrm{C}} + \mu_1^{\mathrm{C}} - \mu^{\mathrm{C}}}{\sigma^{\mathrm{C}}} \\
\Rightarrow \quad & \frac{z^{\mathrm{L}} - \left(\mu_1^{\mathrm{L}} - \mu_1^{\mathrm{C}}\frac{\sigma_1^{\mathrm{L}}}{\sigma_1^{\mathrm{C}}} + \mu^{\mathrm{C}}\frac{\sigma_1^{\mathrm{L}}}{\sigma_1^{\mathrm{C}}}\right)}{\frac{\sigma_1^{\mathrm{L}}\sigma^{\mathrm{C}}}{\sigma_1^{\mathrm{C}}}} \\
\Rightarrow \quad & \mu^{\mathrm{L}} = \mu_1^{\mathrm{L}} - \mu_1^{\mathrm{C}}\frac{\sigma_1^{\mathrm{L}}}{\sigma_1^{\mathrm{C}}} + \mu^{\mathrm{C}}\frac{\sigma_1^{\mathrm{L}}}{\sigma_1^{\mathrm{C}}}, \quad \sigma^{\mathrm{L}} = \frac{\sigma_1^{\mathrm{L}}\sigma^{\mathrm{C}}}{\sigma_1^{\mathrm{C}}}.
\end{aligned} \tag{2}$$

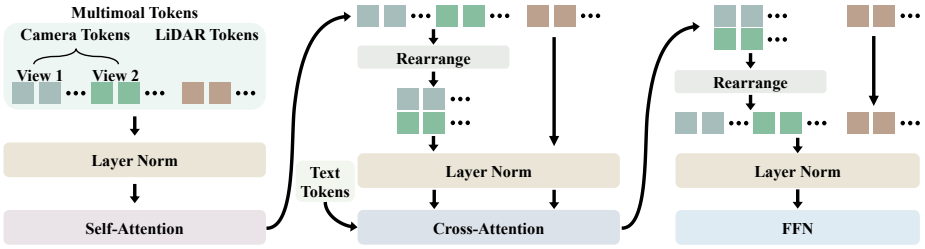

Fig. 2: Transformer block of *UniDriveDreamer*.

3.2 Single-Stage Multimodal World Model

In this section, we present the detailed design of our single-stage multimodal world model, *UniDriveDreamer*. As illustrated in Fig. 1, *UniDriveDreamer* comprises four key components: two modality-specific VAEs, a layout encoder, a text encoder, and a diffusion transformer. Given an input sequence of multi-camera videos $v^{\mathrm{C}} \in \mathbb{R}^{V\times(1+T)\times H^{\mathrm{C}}\times W^{\mathrm{C}}\times 3}$ and LiDAR sweeps $v^{\mathrm{L}} \in \mathbb{R}^{(1+T)\times H^{\mathrm{L}}\times W^{\mathrm{L}}\times 1}$, each modality is first encoded into the latent space $z^{\mathrm{C}} \in \mathbb{R}^{V\times(1+\frac{T}{4})\times\frac{H^{\mathrm{C}}}{8}\times\frac{W^{\mathrm{C}}}{8}\times C}$, $z^{\mathrm{L}} \in \mathbb{R}^{(1+\frac{T}{4})\times\frac{H^{\mathrm{L}}}{8}\times\frac{W^{\mathrm{L}}}{8}\times C}$ by its corresponding VAE. Multi-view textual prompts are encoded using a pretrained text encoder [8]. The layout encoder and the diffusion transformer are described in detail below.

Layout Encoder. The Layout Encoder consists of a spatial downsampling block followed by two spatiotemporal downsampling blocks [36], aligning the spatiotemporal compression rate of layout latents with that of the multimodal data. Specifically, we first project structured 3D scene information, including 3D bounding boxes and HDMap, into both camera views and LiDAR range views. In camera views, different object categories and lane markings are distinguished by distinct colors, while in range views, the projection yields single-channel distance values. To enable the same layout encoder to process both modalities, we replicate the single-channel range-view layout map three times along the channel dimension before feeding it into the encoder. The encoder then outputs multi-camera view condition latents $c^{\mathrm{C}}_{\mathrm{layout}} \in \mathbb{R}^{V\times(1+\frac{T}{4})\times\frac{H^{\mathrm{C}}}{8}\times\frac{W^{\mathrm{C}}}{8}\times C}$ and range-view condition latents $c^{\mathrm{L}}_{\mathrm{layout}} \in \mathbb{R}^{(1+\frac{T}{4})\times\frac{H^{\mathrm{L}}}{8}\times\frac{W^{\mathrm{L}}}{8}\times C}$, which are then used to guide the respective modality synthesis.

Diffusion Transformer. The diffusion transformer architecture comprises three main components: (1) two modality-specific patchifying modules, (2) a shared stack of transformer blocks, and (3) two modality-specific unpatchifying modules. In the patchifying modules, a 3D convolution with a kernel size of (1, 2, 2) followed by a flattening operation is applied to convert the camera-views and range-view inputs into camera tokens T^{C} and LiDAR tokens T^{L} of shapes $\big(B, (V\ L^{\mathrm{C}}), D\big)$ and (B, L^{L}, D), respectively. The camera-view input is constructed by concatenating along the channel dimension the noisy latents $z^{\mathrm{C}}_{\mathrm{t}}$, view-specific embeddings emb^{C}, layout-conditioning latents $c^{\mathrm{C}}_{\mathrm{layout}}$, and first frame

conditioning latents $c^{\mathrm{C}}_{\mathrm{frame}}$; the range-view input is formed analogously. These two sets of modality tokens are then concatenated along the sequence dimension and fed into the shared transformer block stack. As depicted in Fig. 2, the concatenated multimodal tokens first pass through a self-attention module, which jointly models intra-modal spatiotemporal dependencies and enforces cross-modal consistency. The tokens are then separated by modality. The camera tokens are reshaped to $\big((B\ V), L^{\mathrm{C}}, D\big)$ and processed together with the LiDAR tokens through a cross-attention layer that conditions the generation on textual prompts. Subsequently, the camera tokens are reshaped back to their original layout and concatenated with the LiDAR tokens before being passed through a feed-forward network. After passing through the stack of transformer blocks, the modality-specific tokens are individually transformed by their corresponding unpatchifying modules to recover the original latent shape of each modality.

Optimization. *UniDriveDreamer* is optimized by using the flow matching framework [9,24]. Given a timestep $t \in [0, 1]$ sampled from a logit-normal distribution, a random camera noise $z_0^{\mathrm{C}} \in \mathbb{R}^{(1+\frac{T}{4})\times\frac{H^{\mathrm{C}}}{8}\times\frac{W^{\mathrm{C}}}{8}\times C}$, and a random range-view noise $z_0^{\mathrm{L}} \in \mathbb{R}^{(1+\frac{T}{4})\times\frac{H^{\mathrm{L}}}{8}\times\frac{W^{\mathrm{L}}}{8}\times C}$, the noisy latent z_t is obtained as:

$$z_t = t \begin{bmatrix} z_1^{\mathrm{C}} \\ z_1^{\mathrm{L}} \end{bmatrix} + (1-t) \begin{bmatrix} z_0^{\mathrm{C}} \\ z_0^{\mathrm{L}} \end{bmatrix}, \tag{3}$$

where z_1^{C} and z_1^{L} denote the clean camera and LiDAR latents, respectively. The corresponding ground-truth velocity ν_t is computed as:

$$\nu_t = \frac{dz_t}{dt} = \begin{bmatrix} z_1^{\mathrm{C}} - z_0^{\mathrm{C}} \\ z_1^{\mathrm{L}} - z_0^{\mathrm{L}} \end{bmatrix}. \tag{4}$$

The model is then optimized to predict this velocity via the following objective:

$$\mathcal{L} = \mathbb{E}_{z_1, z_0, c, t} \| u(z_t, c, t; \theta) - \nu_t \|^2, \tag{5}$$

where the conditioning signal c comprises layout conditions, multi-view text prompt embeddings, and first frame latents. The parameter set θ includes the weights of the diffusion transformer and the layout encoder.

4 Experiment

In this section, we present our experimental setup, including the datasets, implementation details and evaluation metrics. Subsequently, both quantitative and qualitative results are provided to demonstrate the superior performance of the proposed *UniDriveDreamer*. In addition, ablation studies are conducted to validate key design choices.

4.1 Experiment Setup

Datasets. The training dataset is derived from the nuScenes dataset [7], consisting of 700 training sequences and 150 validation sequences. Each sequence

Table 1: Comparison of the generation quality on nuScenes validation set.

Modality	Method	Camera Generation		LiDAR Generation	
		FID ↓	FVD ↓	MMD ↓	JSD ↓
Camera	DriveDreamer [38]	14.90	340.80	-	-
	DriveDreamer-2 [52]	11.20	55.70	-	-
	MagicDrive [11]	16.20	-	-	-
	MagicDrive-V2 [10]	20.91	94.84	-	-
	Panacea [42]	16.96	139.0	-	-
	Drive-WM [41]	15.80	122.70	-	-
LiDAR	LiDARVAE [6]	-	-	11.0	-
	LiDARGen [55]	-	-	19.0	0.160
	RangeLDM [16]	-	-	1.90	0.054
Multimodal	UniScene [20]	6.12	70.52	1.53	0.072
	OmniGen [33]	21.01	-	2.94	0.105
	Genesis [13]	4.24	16.95	-	-
	UniDriveDreamer	**2.81**	**11.44**	**0.27**	**0.039**

encompasses approximately 20 seconds of recorded driving data, captured by six surround-view cameras and a roof-mounted LiDAR. Following [38,40,52], we preprocess the nuScenes dataset to calculate 12Hz annotations.

Implementation Details. In our experiments, camera videos are processed at a resolution of 432×768 and LiDAR range maps at 32×1024, with a temporal length of 17 frames. *UniDriveDreamer* is initialized with the pretrained weights of Wan-2.1 T2V-1.3B [36]. The model is optimized with a learning rate of 2×10^{-5}. For the LiDAR VAE, we load the pretrained VAE weights from Wan-2.1 [36] and set the learning rate to 5×10^{-5}. The reconstruction loss, KL divergence loss, and LPIPS loss are weighted with coefficients $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = 0.3$ respectively. Furthermore, following [36], we adopt the umT5 model [8] as the text encoder.

Evaluation Metrics. The evaluation in this work covers three key aspects: generation quality, LiDAR reconstruction fidelity, and downstream task performance. To assess the quality of synthesized camera videos, we employ Fréchet Inception Distance (FID) [14] and Fréchet Video Distance (FVD) [35]. For LiDAR sequences, we follow the evaluation protocol of [54] and adopt Maximum Mean Discrepancy (MMD) and Jensen–Shannon Divergence (JSD), where the MMD results are reported scaled by 10^4. The geometric accuracy of LiDAR reconstruction is measured using Chamfer Distance and F-Score as adopted in [33]. Finally, we employ BEVFusion [25] to assess the utility of the generated data for downstream perception and report the nuScenes Detection Score (NDS) and mean Average Precision (mAP) for 3D object detection.

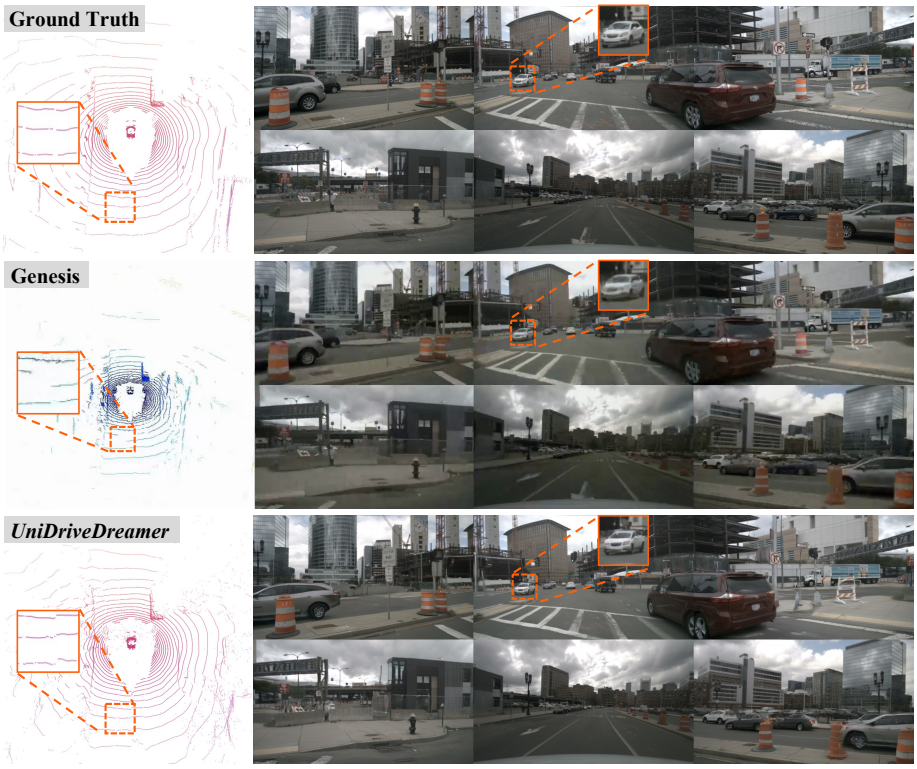

Fig. 3: Qualitative comparison of generated outputs with Genesis [13]. The top row shows the ground truth. The middle row presents results from Genesis. The bottom row displays outputs from our *UniDriveDreamer*.

4.2 Main Results

In this section, we present the experimental results in three aspects: (1) the qualitative and quantitative assessment of multimodal data generation quality; (2) the evaluation of LiDAR reconstruction fidelity; and (3) the analysis of performance gains achieved in downstream perception tasks.

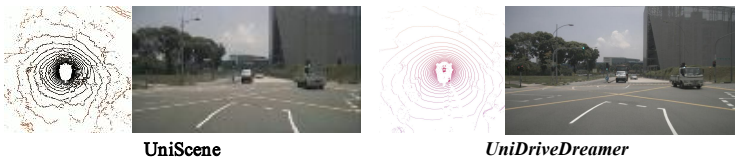

Fig. 4: Qualitative comparison of generated outputs with UniScene [20]. The left side shows results from UniScene, while the right side presents outputs of *UniDriveDreamer*.

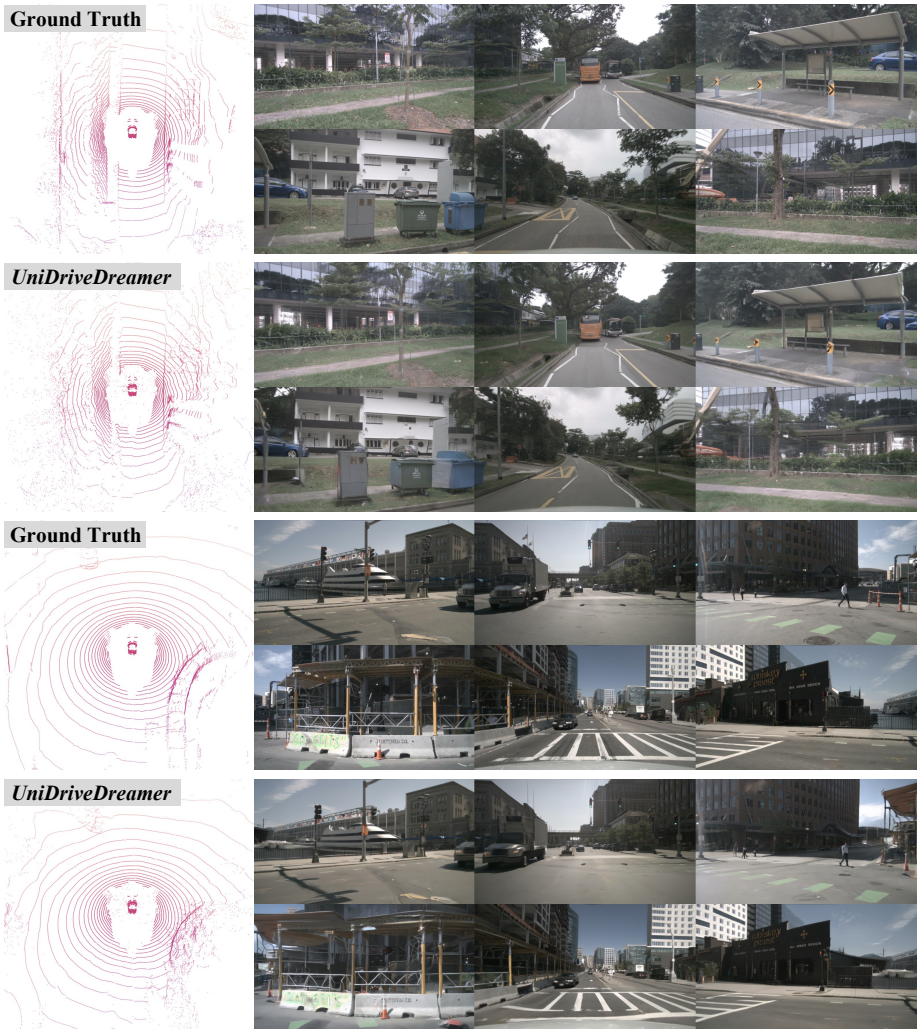

Fig. 5: Qualitative visualization of multimodal outputs generated by *UniDrive-Dreamer*. In each example set, the upper row presents the ground truth, while the lower row shows the corresponding synthesized results generated by *UniDriveDreamer*.

Multimodal Generation Results. As shown in Tab. 1, we compare *UniDrive-Dreamer* against SOTA single-modal and multimodal methods [6, 10, 11, 13, 16, 20, 33, 38, 41, 42, 52, 55] on the nuScenes validation set [7] in terms of generation quality. These quantitative results demonstrate that *UniDriveDreamer* achieves superior performance in both camera video and LiDAR sequence synthesis. Specifically, for video generation, *UniDriveDreamer* attains an FID of 2.81 and an FVD of 11.44. For LiDAR synthesis, it yields an MMD of 0.27 and a JSD of 0.039. These metrics correspond to substantial relative improvements

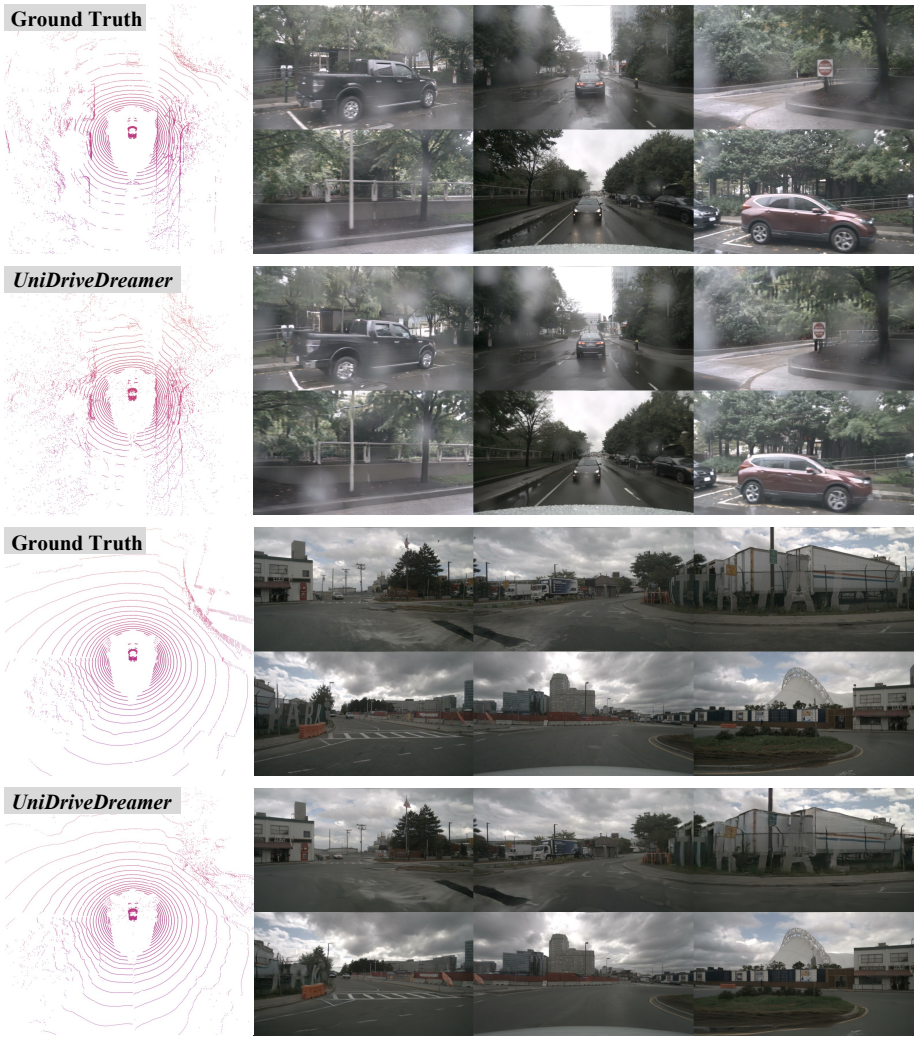

Fig. 6: Qualitative visualization of multimodal outputs generated by *UniDrive-Dreamer*. In each example set, the upper row presents the ground truth, while the lower row shows the corresponding synthesized results generated by *UniDriveDreamer*.

of 82.3% in MMD and 45.8% in JSD when compared to UniScene [20], thereby highlighting the significant advancements in geometric accuracy and distributional alignment achieved by our approach. Qualitative comparisons in Fig. 3 demonstrate the superior generation quality of *UniDriveDreamer*. In the specific context of LiDAR synthesis, the results of *UniDriveDreamer* exhibit more complete ground coverage and higher geometric fidelity than those of Genesis [13], showing closer alignment with the ground truth. Concurrently, regarding cam-

Table 2: Comparison of the reconstruction quality on nuScenes validation set.

Method	Chamfer Distance ↓	F-Score ↑
OmniGen [33]	0.793	0.742
UniDriveDreamer	**0.154**	**0.900**

Table 3: Multimodal generation data augmentation for 3D object detection.

Training Data	mAP ↑	NDS ↑
Original Data	66.38	70.01
Augmented Data by *UniDriveDreamer*	**67.18**	**70.53**

era video generation, our method produces visually sharper and more realistic outputs with enhanced detail preservation. Beyond the comparison with Genesis, we further compare our method with UniScene in Fig. 4. The results show that *UniDriveDreamer* produces LiDAR generations with stronger structural fidelity and fewer noise artifacts, while also achieving higher image quality. Additional generated examples are presented in Fig. 5 and Fig. 6. These visual results further demonstrate the high controllability of our model under varied scene layouts, its robust spatiotemporal consistency across extended sequences, and its precise cross-modal alignment between camera views and LiDAR sweeps.

LiDAR Reconstruction Results. As shown in Tab. 2, we conduct an evaluation of our LiDAR VAE against OmniGen [33] regarding LiDAR sequence reconstruction capabilities. The quantitative results demonstrate that our LiDAR VAE substantially outperforms the current SOTA in terms of reconstruction fidelity, achieving a Chamfer Distance of 0.154 and an F-Score of 0.900. This corresponds to an improvement of approximately 80.6% in Chamfer Distance and 21.3% in F-Score over the baseline. Fig. 7 provides a visual comparison across several distinct driving scenarios between our reconstructed LiDAR sweeps and the ground truth. The visualization clearly demonstrates that the proposed LiDAR VAE preserves geometric structures effectively, thereby generating dense and spatially coherent point clouds that exhibit a high degree of fidelity to the original scans.

Downstream Task Results. Given that the synthesized data inherently lacks LiDAR intensity information, we standardize the intensity channel to zero for all data points during the downstream task training phase. Leveraging these preprocessed real data, we train BEVFusion [25] from scratch to establish our baseline. As shown in Tab. 3, incorporating the data generated by *UniDriveDreamer* yields notable enhancements in downstream perception performance relative to the baseline. Specifically, our approach increases the mAP by +0.8 and the NDS by +0.52 when compared to training exclusively with real data. These results not only validate the effectiveness of *UniDriveDreamer* in enhancing perception performance for autonomous driving but also provide compelling

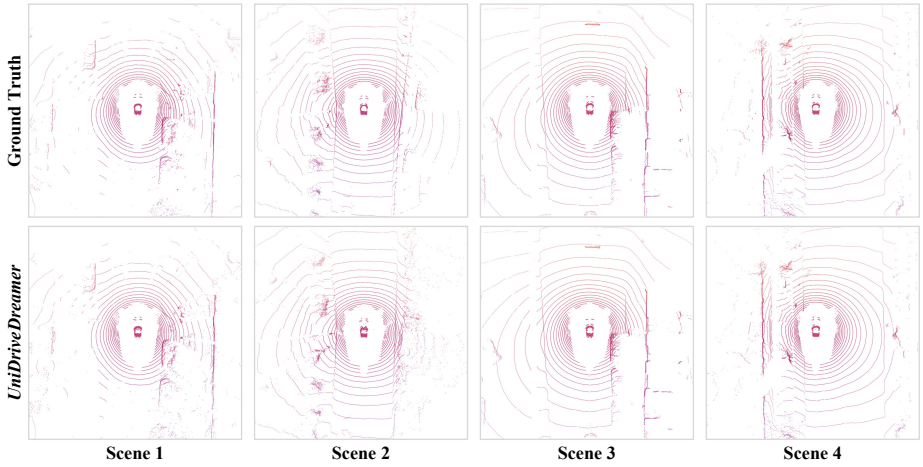

Fig. 7: Qualitative visualization of LiDAR reconstruction. The upper row shows the real LiDAR data, while the lower row presents the reconstructed point clouds from the LiDAR-specific VAE in *UniDriveDreamer*.

Table 4: Effect of range-map repeat height for LiDAR reconstruction fidelity.

Repeat Height	Chamfer Distance ↓	F-Score ↑
1	0.513	0.794
2	0.336	0.829
4	**0.154**	**0.900**

indirect evidence of the high fidelity of its synthesized data and the robustness of its cross-modal consistency.

4.3 Ablation Study

In this section, we conduct ablation studies to analyze two key design choices. First, we investigate the specific effect of range-map height repetition on the fidelity of LiDAR reconstruction.Second, we evaluate the critical impact of Unified Latent Anchoring (ULA) on both cross-modal alignment precision and the overall quality of multimodal generation.

Effect of Range-Map Repeat Height. We investigate the influence of height repetition in range-map encoding on LiDAR reconstruction quality. As reported in Tab. 4, progressively increasing the number of height repetitions from 1 to 4 leads to consistent improvements in Chamfer Distance and F-Score. The best performance is achieved with 4 repetitions, reaching a Chamfer Distance of 0.154 and an F-Score of 0.900, which validates that repeated height channels effectively mitigate range-map sparsity and improve geometric detail recovery. In light of these empirical findings, we formally adopt 4 repetitions as the default configuration for our proposed model.

Fig. 8: Qualitative comparison of generated outputs with and without the Unified Latent Anchoring (ULA) module. The top row displays the ground truth. The middle row illustrates results without ULA, highlighting geometric inconsistencies and cross-modal misalignment between LiDAR points and RGB content. The bottom row presents results with ULA,showing enhanced cross-modal geometric coherence and improved visual quality.

Effect of ULA. We ablate the proposed Unified Latent Anchoring (ULA) module to examine its contribution to training stability and cross-modal alignment. ULA is designed to unify the mathematical distributions of RGB and LiDAR latent representations without introducing any additional learnable parameters, thereby facilitating more effective joint multimodal training. It is worth noting that ULA does not directly impose semantic alignment across modalities; instead, the semantic correspondence is subsequently learned by the Transformer on top of the unified latent space. As illustrated in Fig. 8, removing ULA leads to a pronounced degradation in the consistency between generated RGB images and LiDAR outputs, indicating that explicit latent distribution alignment is critical for improving intra-modal generation quality and maintaining robust cross-modal coherence. Furthermore, as shown in Fig. 9, ULA accelerates model convergence, with the training loss approaching its converged value more rapidly.

Fig. 9: Comparison of training loss curves, including both raw data and smoothed curves. The blue curves denote the model trained without ULA, while the red curves denote the model trained with ULA.

We further evaluate the cross-dataset robustness of ULA on the Waymo dataset. Without ULA, the model obtains 8.62 FID and 1.54 MMD. When using ULA estimated from nuScenes statistics, the performance improves to 3.58 FID and 0.38 MMD, and further improves to 2.13 FID and 0.22 MMD when using Waymo-estimated ULA statistics. These results demonstrate that ULA remains effective across datasets and validate its importance in enabling stable and robust joint multimodal learning.

5 Conclusion

This paper introduces *UniDriveDreamer*, a single-stage unified multimodal world model. To effectively bridge the inherent modality gap between camera and LiDAR data, our approach adopts range images as the canonical LiDAR representation and incorporates a specifically tailored LiDAR VAE. Furthermore, we propose Unified Latent Anchoring (ULA), a mechanism dedicated to explicit cross-modal distribution alignment, which ensures semantic consistency across heterogeneous sensor inputs. Built upon a shared diffusion transformer, the model directly synthesizes temporally consistent and geometrically coherent multi-camera videos and LiDAR sweeps. Notably, this end-to-end generation process eliminates the reliance on cascaded pipelines or intermediate representations, thereby streamlining the synthesis workflow. Extensive experiments validate that *UniDriveDreamer* surpasses SOTA methods in both camera and LiDAR synthesis. In the domain of video generation, our model achieves an FID of 2.81 and an FVD of 11.44. Regarding LiDAR generation, it attains an MMD of 0.27 and a JSD of 0.039. These metrics correspond to substantial relative improvements of 82.3% in MMD and 45.8% in JSD when compared to the best prior work, highlighting superior geometric accuracy. Moreover, downstream perception tasks benefit measurably from the synthesized multimodal data, with improvements of +0.8 mAP and +0.52 NDS, confirming its practical utility for enhancing autonomous driving systems. Our work provides an efficient paradigm for multimodal sensor simulation, paving the way for more data-efficient and robust autonomous driving development.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Alhaija, H.A., Alvarez, J., Bala, M., Cai, T., Cao, T., Cha, L., Chen, J., Chen, M., Ferroni, F., Fidler, S., et al.: Cosmos-transfer1: Conditional world generation with adaptive multimodal control. arXiv preprint arXiv:2503.14492 (2025)
3. Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)
4. Ball, P.J., Bauer, J., Belletti, F., Brownfield, B., Ephrat, A., Fruchter, S., et al.: Genie 3: A new frontier for world models (2025), `https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/`, accessed: 2026-06-28
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Caccia, L., Van Hoof, H., Courville, A., Pineau, J.: Deep generative modeling of lidar data. In: IROS (2019)
7. 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: CVPR (2020)
8. Chung, H.W., Garcia, X., Roberts, A., Tay, Y., Firat, O., Narang, S., Constant, N.: Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In: ICLR (2023)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
10. 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. In: CVPR (2025)
11. Gao, R., Chen, K., Xie, E., HONG, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. In: ICLR (2023)
12. 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. NeurIPS (2024)
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. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)
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, Q., Zhang, Z., Hu, W.: Rangeldm: Fast realistic lidar point cloud generation. In: ECCV (2024)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

18. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
19. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review (2022)
20. 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: CVPR (2025)
21. Li, X., Zhang, Y., Ye, X.: Drivingdiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model. In: ECCV (2024)
22. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE TPAMI (2024)
23. Liang, D., Zhang, D., Zhou, X., Tu, S., Feng, T., Li, X., Zhang, Y., Du, M., Tan, X., Bai, X.: Seeing the future, perceiving the future: A unified driving world model for future generation and perception. arXiv preprint arXiv:2503.13587 (2025)
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
25. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: ICRA (2023)
26. Ma, E., Zhou, L., Tang, T., Zhang, Z., Han, D., Jiang, J., Zhan, K., Jia, P., Lang, X., Sun, H., et al.: Unleashing generalization of end-to-end autonomous driving with controllable long video generation. arXiv preprint arXiv:2406.01349 (2024)
27. Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Chen, X., Jia, G., Huang, G., Mei, W.: Recondreamer-rl: Enhancing reinforcement learning via diffusion-based scene reconstruction. arXiv preprint arXiv:2508.08170 (2025)
28. Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Huang, G., Liu, C., Chen, Y., Wang, Y., Zhang, X., et al.: Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In: CVPR (2025)
29. Ren, X., Lu, Y., Cao, T., Gao, R., Huang, S., Sabour, A., Shen, T., Pfaff, T., Wu, J.Z., Chen, R., et al.: Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. arXiv preprint arXiv:2506.09042 (2025)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
31. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
32. Swerdlow, A., Xu, R., Zhou, B.: Street-view image generation from a bird's-eye view layout. IEEE Robotics and Automation Letters (2024)
33. Tang, T., Ma, E., Zhou, X., Wang, L., Yan, T., Zhang, X., Zhan, K., Jia, P., Lang, X., Bian, J.W., et al.: Omnigen: Unified multimodal sensor generation for autonomous driving. In: ACM MM (2025)
34. Team, G., Ye, A., Wang, B., Ni, C., Huang, G., Zhao, G., Li, H., Zhu, J., Li, K., Xu, M., et al.: Gigaworld-0: World models as data engine to empower embodied ai. arXiv preprint arXiv:2511.19861 (2025)
35. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)

36. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
37. Wang, W., Zhu, J., Zhang, Z., Wang, X., Zhu, Z., Zhao, G., Ni, C., Wang, H., Huang, G., Chen, X., et al.: Drivegen3d: Boosting feed-forward driving scene generation with efficient video diffusion. arXiv preprint arXiv:2510.15264 (2025)
38. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: ECCV (2024)
39. Wang, X., Zhu, Z., Huang, G., Wang, B., Chen, X., Lu, J.: Worlddreamer: Towards general world models for video generation via predicting masked tokens. arXiv preprint arXiv:2401.09985 (2024)
40. Wang, X., Zhu, Z., Zhang, Y., Huang, G., Ye, Y., Xu, W., Chen, Z., Wang, X.: Are we ready for vision-centric driving streaming perception? the asap benchmark. In: CVPR (2023)
41. 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: CVPR (2024)
42. 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)
43. Yan, H., Zhong, Z., Zhu, J., He, J., Yuan, W., Song, W., Gong, X., Cai, Y., Zhao, G., Yan, X., et al.: S-vam: Shortcut video-action model by self-distilling geometric and semantic foresight. arXiv preprint arXiv:2603.16195 (2026)
44. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., et al.: Generalized predictive model for autonomous driving. In: CVPR (2024)
45. 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)
46. Yang, Z., Guo, X., Ding, C., Wang, C., Wu, W., Zhang, Y.: Instadrive: Instance-aware driving world models for realistic and consistent video generation. In: ICCV (2025)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
49. 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. arXiv preprint arXiv:2407.05679 (2024)
50. 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: CVPR (2025)
51. Zhao, G., Wang, X., Ni, C., Zhu, Z., Qin, W., Huang, G., Wang, X.: Recondreamer++: Harmonizing generative and reconstructive models for driving scene representation. In: ICCV (2025)
52. 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. In: AAAI (2025)
53. Zhu, Z., Wang, X., Zhao, W., Min, C., Li, B., Deng, N., Dou, M., Wang, Y., Shi, B., Wang, K., et al.: Is sora a world simulator? a comprehensive survey on general world models and beyond. arXiv preprint arXiv:2405.03520 (2024)

54. Zyrianov, V., Che, H., Liu, Z., Wang, S.: Lidardm: Generative lidar simulation in a generated world. In: ICRA (2025)
55. Zyrianov, V., Zhu, X., Wang, S.: Learning to generate realistic lidar point clouds. In: ECCV (2022)