LiSTAR: Ray-Centric World Models for 4D LiDAR Sequences in Autonomous Driving

Pei Liu[1,2*], Songtao Wang[3*], Lang Zhang[3*†], Xingyue Peng[3], Yuandong Lyu[3], Jiaxin Deng[3], Songxin Lu[3], Weiliang Ma[3], Xueyang Zhang[3], Yifei Zhan[3], Kun Zhan[3], and Jun Ma[1,2‡]

[1] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2] The Hong Kong University of Science and Technology, Hong Kong SAR, China
[3] Li Auto Inc., Beijing, China
jun.ma@ust.hk

Abstract. Synthesizing high-fidelity and controllable 4D LiDAR data is crucial for creating scalable simulation environments for autonomous driving. This task is inherently challenging due to the sensor's unique spherical geometry, the temporal sparsity of point clouds, and the complexity of dynamic scenes. To address these challenges, we present LiSTAR, a novel generative world model that operates directly on the sensor's native geometry. LiSTAR introduces a Hybrid-Cylindrical-Spherical (HCS) representation to preserve data fidelity by mitigating quantization artifacts common in Cartesian grids. To capture complex dynamics from sparse temporal data, it utilizes a Spatio-Temporal Attention with Ray-Centric Transformer (START) that explicitly models feature evolution along individual sensor rays for robust temporal coherence. Furthermore, for controllable synthesis, we propose a novel 4D point cloud-aligned voxel layout for conditioning and a corresponding discrete Masked Generative START (MaskSTART) framework, which learns a compact, tokenized representation of the scene, enabling efficient, high-resolution, and layout-guided compositional generation. Comprehensive experiments validate LiSTAR's state-of-the-art performance across 4D LiDAR reconstruction, prediction, and conditional generation. Our method achieves substantial quantitative gains: improving reconstruction IoU by up to 85%, lowering prediction L1 Med by up to 64%, and reducing generation MMD by 61%. This level of performance provides a powerful new foundation for creating realistic and controllable simulations for autonomous driving systems. Project Page: https://ocean-luna.github.io/LiSTAR.github.io/.

Keywords: World models, Controllable generation, 4D LiDAR generation, Point cloud processing

---

* These authors contributed equally to this work.
† Project Leader.
‡ Corresponding author.

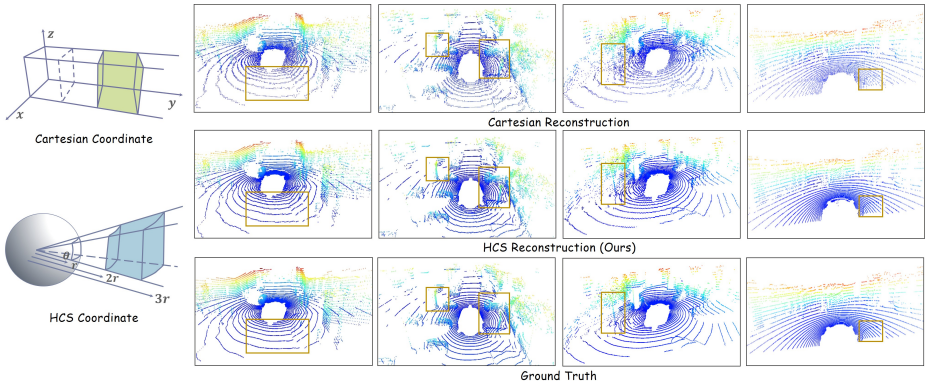

Fig. 1: Cartesian vs. HCS coordinate for LiDAR scene representation. Cartesian coordinate partitions space into uniform, axis-aligned cubes, ignoring the native ray geometry of LiDAR. HCS coordinates divides space into angular–radial cells centered at the sensor origin, aligning with LiDAR's ray-based sampling pattern and preserving range-dependent resolution.

1 Introduction

World models, which aim to internalize environmental dynamics by learning generative predictors, have demonstrated strong capabilities across a wide range of visual and interactive tasks and are now increasingly explored for autonomous driving [17, 35]. Recent progress has largely focused on structured modalities like videos and occupancy grids, whose dense organization fits well with established processing pipelines [11, 52]. By contrast, LiDAR remains understudied despite its importance for accurate 3D geometry and all-weather perception. The sparse, unordered, and irregular nature of LiDAR point clouds [22, 31, 58] poses fundamental challenges for generative modeling, limiting the direct adoption of techniques designed for regularly structured data.

Despite recent progress in LiDAR scene synthesis [26, 62, 70], significant hurdles remain. A primary challenge stems from conventional voxelization, which converts LiDAR returns into dense Cartesian grids. As illustrated in Fig. 1, this approach overlooks the native ray-based sampling geometry of spinning sensors, leading to quantization artifacts and distorted structural patterns that impair fidelity [24, 27, 64]. Furthermore, the inherent sparsity and non-uniform sampling of point clouds complicate the preservation of temporal coherence, often resulting in flickering surfaces or inconsistent dynamic object alignment [19]. Finally, for controllable synthesis, the prevalent reliance on temporal Bird's-Eye-View (BEV) layouts [40, 41] as conditional inputs imposes a critical bottleneck. This 2D projection inherently flattens the rich 3D world, constraining the ability to precisely guide generation or manipulate objects in full 3D space, a capability crucial for targeted scenario design and safety evaluation.

To address these challenges, we introduce LiSTAR, a novel world model built upon a pioneering Hybrid-Cylindrical-Spherical (HCS)-based 4D Vector Quantised-Variational AutoEncoder (VQ-VAE) [3, 51] to learn a discrete representation of LiDAR scenes. LiSTAR begins with a novel Hybrid-Cylindrical-Spherical (HCS) volumetric representation. Unlike conventional range images [38, 40], which project 3D points onto a 2.5D plane, inevitably causing information loss due to occlusion and quantization errors in complex scenes, our HCS scheme discretizes the native sensor space into a dense 3D grid. This volumetric approach preserves the complete geometric topology of LiDAR scans, including multiple returns along a single ray, thereby offering a richer and more faithful substrate for generative modeling than flattened projections. Building on this, we introduce the Spatio-Temporal Attention with Ray-Centric Transformer (START). While standard transformers treat voxels as unordered tokens or apply generic window attention, START incorporates a strong inductive bias aligned with LiDAR physics. It explicitly models feature correlations along the sensor's ray direction through Ray-Centric Attention, capturing the unique radial dependencies of spinning LiDARs, while simultaneously enforcing causal temporal consistency. Finally, to enable controllable synthesis, we propose MaskSTART. Distinct from the standard MaskGIT [5], which typically operates on 2D images, MaskSTART is tailored for 4D sequence modeling. It introduces a dual-conditioning mechanism that integrates historical context with a novel 4D point cloud-aligned voxel layout. This allows for precise, geometry-aware control over the generation process, enabling the synthesis of complex, temporally coherent driving scenarios that strictly adhere to user-defined structural constraints.

We conduct extensive experiments on the large-scale nuScenes benchmark, evaluating LiSTAR on a suite of tasks including point cloud reconstruction, prediction, and generation. In both unconditional and layout-conditioned settings, LiSTAR consistently outperforms state-of-the-art baselines. Beyond quantitative metrics, we demonstrate that the framework's ability to produce controllable, temporally consistent LiDAR sequences unlocks novel downstream applications. The main contributions of this work are:

- We present LiSTAR, a 4D LiDAR world model that unifies HCS representation, START, and MaskSTART into a single end-to-end framework, explicitly tailored to LiDAR's acquisition geometry and temporal dynamics for world models of autonomous driving.
- We propose an HCS coordinate voxelization scheme that preserves the native ray structure and range resolution, effectively mitigating geometric distortion caused by conventional Cartesian discretization.
- We design the START module, which models feature correlations along LiDAR rays to capture spatial structure and temporal dependencies jointly, ensuring geometric fidelity and frame-to-frame consistency.
- We introduce a MaskSTART pipeline for 4D LiDAR sequences that supports fine-grained semantic conditioning on 4D point cloud-aligned voxel layout. This approach enables controllable and diverse scenario synthesis, allowing for precise manipulation and generation of complex scene structures.

– We achieve state-of-the-art performance on large-scale autonomous driving benchmarks for both point cloud reconstruction, prediction, and generation, and demonstrate LiSTAR's utility in realistic, controllable simulation scenarios.

2 Related Work

2.1 3D Representation for Point Clouds

Choosing an effective 3D representation is critical for point cloud generation. Point-based approaches, such as PointNet and PointNet++ [29, 43, 45], directly operate on raw points, aggregating local and global features to encode spatial context. Voxelization [33, 44, 59] discretizes the space into dense grids but is memory-intensive, leading to sparse convolution designs [7, 14] that skip empty cells. Projection-based representations are also popular: BEV [23, 59] vertically projects points onto a planar map, while range images [18, 38, 40, 46, 70] map points to polar coordinates to form 2.5D grids. Recent works leverage VAE-family models [10, 21, 51] for latent compression, e.g., VQ-VAE [3, 51] learns discrete codebooks for compact feature tokens. For simulation, ray-casting pipelines [2, 9, 34, 39] reproduce ray-drop patterns from virtual assets. Finally, implicit neural representations, such as NeRF [37], allow differentiable rendering of point clouds from learned occupancy or semantic fields [24, 63].

2.2 World Models for Point Clouds

World models predict future observations from historical states and agent actions, enabling agents to model temporal dynamics. While early work focused on image/video prediction [17, 48, 56, 65, 66], recent studies extend to structured 3D data. In 3D occupancy world models [15, 25, 53, 67, 68], discrete volumetric tokens are predicted to maintain spatial consistency. Point cloud world models [60, 64, 69] predict temporal LiDAR sequences by combining latent tokenization and generative backbones. For instance, Copilot4D [64] encodes LiDAR frames with VQ-VAE and applies discrete diffusion for forecasting, while LiDAR-DM [69] adapts diffusion transformers for long-horizon predictions. These approaches highlight the promise and current limitations of scalable, geometry-aware token-based generative modeling for 4D LiDAR.

2.3 Diffusion Models

Diffusion models learn a forward process that progressively corrupts data with noise, and a reverse process to recover the original signal. For continuous data, Gaussian perturbations [16, 47, 49, 50] are widely used due to favorable statistical properties enabling stable training objectives. MaskGIT [5] replaces Gaussian noise with aggressive token masking and BERT-style training [8], outperforming Gaussian diffusion in several domains, including video [61] and point clouds [57].

Beyond diffusion, flow matching [30] learns continuous-time flows between base and target distributions, allowing ODE-based sampling with convergence guarantees, and has been applied to sequences [13] and spatial modalities. Recent works adapt these generative paradigms to 3D point clouds, using token-based pipelines, view-consistent constraints, and geometry-aware noise schedules to improve spatial fidelity and temporal stability, making them promising backbones for LiDAR world modeling.

3 Methodology

We introduce LiSTAR, a novel generative world model for 4D LiDAR synthesis, composed of two synergistic components: an HCS-based 4D VQ-VAE for representation learning and a MaskSTART model for prediction and generation.

The HCS-based 4D VQ-VAE, shown in Fig. 2 (left), first transforms the input LiDAR sequence into a compact, discrete latent space. The encoder employs stacked START blocks, featuring Spatial Ray-Centric Attention (SRA) and Cyclic-Shifted Temporal Causal Attention (CSTA), to capture spatio-temporal dynamics and produce a quantized codebook representation effectively. This discrete representation then serves as the foundation for the MaskSTART model (Fig. 2, right), a unified framework that performs masked generative modeling for both prediction and conditional generation. In the generation task, it conditions on 4D point cloud-aligned voxel layouts, which are fused via a zero-initialized adapter to guide the synthesis of realistic and semantically consistent sequences. Further details on the algorithmic procedures for reconstruction, prediction, and generation are provided in the Supplementary Material.

3.1 HCS Coordinate Voxelization

Conventional Cartesian voxelization forces a trade-off between fidelity and efficiency: high-resolution grids needed for detail are massively sparse and computationally expensive. We overcome this by proposing a voxelization scheme in an HCS Coordinate System that mirrors the native spherical projection of LiDAR sensors, as shown in Fig. 1. Our method partitions space into bins of constant angular resolution, preserving geometric details at all ranges while yielding a compact and efficient representation.

Formally, a point cloud is defined as:

$$P = \{\mathbf{p}^{(n)} \in \mathbb{R}^3 \;\mid\; 1 \le n \le N\},$$

where each point $\mathbf{p}^{(n)} = (x^{(n)}, y^{(n)}, z^{(n)})$ is expressed in Cartesian coordinate. We map these points into the HCS coordinate system $(\rho^{(n)}, \theta^{(n)}, \phi^{(n)})$ using:

$$\begin{cases} \rho^{(n)} = \sqrt{(x^{(n)})^2 + (y^{(n)})^2}, \\ \theta^{(n)} = \arctan 2(y^{(n)}, x^{(n)}), \\ \phi^{(n)} = \arctan 2(z^{(n)}, \rho^{(n)}). \end{cases} \tag{1}$$

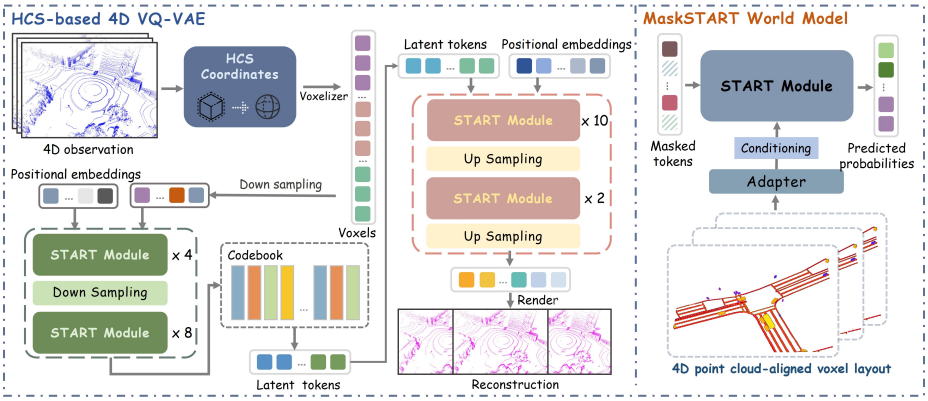

Fig. 2: Illustration of the LiSTAR framework for 4D LiDAR sequence reconstruction and generation. The framework begins by voxelizing LiDAR point clouds into a spherical coordinate representation, which is downsampled and processed by multiple START modules in the encoder to extract semantic-rich latent tokens. The decoder reconstructs detailed 4D sequences by up-sampling tokens with additional START modules. The MaskSTART component facilitates controllable and diverse generation by predicting masked tokens using a bidirectional transformer, conditioned on 4D point cloud-aligned voxel layouts. This design captures spatiotemporal dependencies while preserving fine-grained geometric details.

The voxelization is performed by a function $\mathrm{V} : (\rho, \theta, \phi) \rightarrow (i, j, k)$. Crucially, distinct from 2.5D range maps that store only the nearest depth value per pixel (j, k), our HCS representation maintains a full 3D occupancy grid $\mathcal{G} \in \{0, 1\}^{R \times H \times W}$. This allows LiSTAR to explicitly model the sparsity and distribution of points along the ray dimension R, effectively capturing complex structures that are otherwise compressed in projection-based methods.

3.2 START Module

We introduce the START module, a novel 4D attention mechanism specifically designed for sequential LiDAR data. START operates in a causal temporal manner to capture motion patterns while leveraging spatial ray-centric attention, explicitly aligned with the intrinsic geometry of LiDAR sensors. This formulation enables the model to capture both spatiotemporal dependencies and fine-grained structural relationships effectively, as illustrated in Fig. 3.

Spatial Ray-Centric Attention To explicitly encode the intrinsic ray-like structure of LiDAR scans, we introduce the Ray-Centric Attention (RA) layer. It operates on a dense tensor $V_{\rho\theta\phi} \in \mathbb{R}^{l \times h \times w}$, where l, h, and w correspond to the ray, vertical, and horizontal angular dimensions, respectively. To efficiently process this representation, we first unfold (F) the tensor V along its ray dimension l, transforming it into a 2D matrix of size $d \times l$, where $d = h \times w$. A

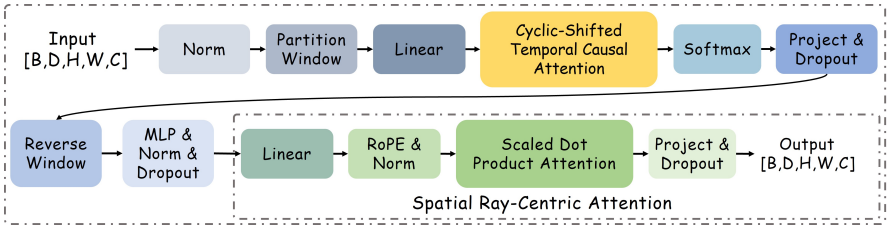

Fig. 3: An illustration of our START module. It processes a 4D feature map of shape $[B, D, H, W, C]$, where D is the temporal dimension. It is composed of two key components: (1) a CSTA block that operates on windowed features to efficiently model temporal dependencies, and (2) an SRA block that processes features reshaped to $[B * D, H, W, C]$ to capture spatial correlations along the ray dimension.

standard self-attention mechanism is then applied to a normalized version of this flattened representation, denoted V_n. The output is computed as:

$$V_n = \mathrm{Norm}(\mathrm{F}(V)), \tag{2}$$

$$V' = softmax\left(\frac{V_n W_Q (V_n W_K)^T}{\sqrt{d_k}}\right) V_n W_v, \tag{3}$$

where W_Q, W_K, and W_V are linear projection matrices that map the input features into query, key, and value spaces. This mechanism enables each ray to aggregate information from all other rays based on their learned similarities, effectively capturing the global context of the 3D scene.

Spatial RA (SRA) block embeds the proposed RA layer within a pre-norm Transformer architecture. The process involves a residual connection after the RA-layer, followed by Layer Normalization (LN) and a Feed-Forward Network (FFN) with a second residual connection. Formally, this is expressed as:

$$V_{\mathrm{SRA}} = \mathrm{F}^{-1}\left((V' + f(V')) + g(\mathrm{LN}(V' + f(V')))\right), \tag{4}$$

where $f(\cdot)$ and $g(\cdot)$ represent the RA layer and the FFN, respectively. F^{-1} indicates the inverse function of F. This residual-in-residual design ensures stable training while enhancing feature expressiveness.

Applying standard self-attention to the full 4D voxel grid is not only computationally prohibitive but also structurally agnostic to the underlying LiDAR geometry. It fails to exploit the inherent radial dependencies captured by sensor rays. Our SRA addresses this by restricting attention computations to the radial dimension. This targeted approach allows the model to efficiently reason about occlusions and spatial relationships along each line of sight. Consequently, SRA captures fine-grained local structures while maintaining global context, all within a feasible computational budget.

Cyclic-Shifted Temporal Causal Attention To effectively model 4D LiDAR sequences, our model must address two fundamental challenges: the spatial discontinuity arising from spherical coordinate projection, and the need for causal temporal modeling. To this end, we propose the CSTA module, which is composed of two specialized mechanisms. First, Cyclic-Shifted Window Attention (CSWA) restores spatial continuity across the azimuthal seam by leveraging a shifted-windowing scheme inspired by the Swin Transformer [32]. Second, Temporal Causal Attention (TCA) enforces a strict chronological order to learn valid motion patterns without information leakage from the future.

Cyclic-Shifted Window Attention. The projection of LiDAR's HCS coordinates onto a discrete grid introduces a critical spatial discontinuity. Points that are physically adjacent across the azimuthal seam ($0°/360°$) are placed at opposite ends of the tensor's width dimension, artificially severing their geometric relationship.

This issue can be conceptualized by considering the mapping of a local neighborhood. Let $\mathcal{N}(p)$ denote a connected 3D neighborhood of a point p located on the azimuthal boundary. The projection onto a tensor of width W splits this single neighborhood into two disjoint sets of indices at the extremes of the tensor dimension:

$$\mathcal{N}(p) \xrightarrow{\text{Projection}} \cdots, w_{W-2}, w_{W-1} \cup w_0, w_1, \cdots. \tag{5}$$

Consequently, standard network operators with fixed receptive fields (e.g., convolutions, window attention) fail to process this neighborhood cohesively. They perceive the two parts as maximally distant, leading to feature artifacts and an incomplete understanding of the global scene structure.

To address the boundary discontinuity induced by spherical coordinate unwrapping, we propose CSWA. CSWA explicitly models the periodic nature of the azimuthal dimension, enabling information flow across the artificial seam. The mechanism operates in two alternating stages. A standard Window MSA (W-MSA) first computes self-attention within local, non-overlapping windows for efficient feature extraction. Subsequently, a Shifted Window MSA (SW-MSA) block introduces a cyclic shift along the azimuthal dimension. This realigns the window grid, enabling cross-window connections, particularly across the $0°/360°$ seam. A carefully designed attention mask ensures that interactions are confined to valid local regions in the shifted configuration before the shift is reversed.

By alternating these standard and shifted window configurations, CSWA achieves a global receptive field with linear complexity, efficiently restoring the topological continuity of the spherical space.

Temporal Causal Attention. Beyond static spatial features, modeling temporal dynamics is crucial for interpreting motion and ensuring coherence across LiDAR frames. Standard attention mechanisms are permutation-invariant and thus non-causal, allowing information to leak from future frames, which is invalid for predictive tasks.

To address this, we introduce TCA, a mechanism designed to model scene evolution while strictly adhering to the arrow of time. TCA extends the causal

constraint to a history of L preceding frames, $\{X_{t-L}, X_{t-L+1}, \cdots, X_{t-1}\}$, allowing the model to capture long-range temporal dependencies. Queries Q_t are generated from the current frame X_t, while a unified set of keys K_{hist} and values V_{hist} is created by concatenating the respective projections from all L past frames. The attention mechanism then aggregates information from the entire history as follows:

$$\mathrm{TCA}(X_t) = \mathrm{softmax}\left(\frac{Q_t K_{\mathrm{hist}}^T}{\sqrt{d_k}}\right) V_{\mathrm{hist}}. \tag{6}$$

This formulation ensures that the model's output for time t depends on past and present information, enabling it to learn robust and complex motion patterns from a rich temporal context.

We integrate TCA by interleaving it with CSWA layers. This layered structure allows the model to jointly refine spatial details and update temporal states, creating a comprehensive 4D representation. This unified approach is essential for tasks requiring both spatial integrity and temporal coherence, such as dynamic scene reconstruction and motion forecasting.

3.3 Training Objectives

The encoder-decoder framework is trained by minimizing a loss function with three components: vector quantization loss, voxel reconstruction loss, and point cloud reconstruction loss:

$$L = L_{VQ}(z, \hat{z}) + L_v(V_{pred}, V_{target}) + L_p(U_{pred}, V_{target}), \tag{7}$$

where z and $\hat{z}$ are encoder outputs and quantized features, respectively; V_{pred} denotes the predicted voxels, U_{pred} represents voxelized rendered point cloud; and V_{target} denotes target voxels. This loss formulation ensures fidelity across both voxel representation and rendered point clouds, enhancing reconstruction quality.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. Our experiments are conducted on three widely adopted autonomous driving datasets: the large-scale nuScenes dataset [4], KITTI Odometry [12], and KITTI-360 [28]. While nuScenes provides dense, 360-degree point clouds from a 32-beam LiDAR, KITTI and KITTI-360 capture high-resolution scenes using 64-beam LiDAR sensors. **Data Representation.** We voxelize point clouds into a high-resolution HCS grid to preserve fine-grained geometric details. Specifically, we set the resolution to $(640 \times 1088 \times 32)$ for nuScenes and $(640 \times 1088 \times 64)$ for KITTI and KITTI-360. The model processes temporal sequences with a total duration of 4 s, allowing for long-term dependency modeling. For the prediction

task, we define the operational range as [-70, 70]m in x/y and [-4.5, 4.5]m in z, while for the generation task, the range is [-50, 50]m in x/y and [-3, 5]m in z. **Network Architecture.** The VQ-VAE encoder is composed of 12 stacked START blocks with a hidden dimension of 512. The codebook size is set to 2048 with a dimension of 1024. The MaskSTART is deepened to 24 layers with 8 attention heads and a hidden dimension of 1280 to capture complex spatio-temporal dynamics. **Training & Inference.** We train the VQ-VAE for 100 epochs and MaskSTART for 200 epochs using the AdamW optimizer with a learning rate of $5e-5$ with a batch size of 2. During inference, we set the CFG to 1.0.

Table 1: Point cloud reconstruction results on nuScenes and KITTI-360. LiSTAR significantly outperforms the baseline across all metrics. Notably, it achieves up to an 85% relative increase in IoU and a 79% reduction in MMD, demonstrating superior geometric accuracy and distribution similarity. Best results are in bold. (↑: Higher is better, ↓: Lower is better).

Dataset	Method	IoU↑	Chamfer↓	MMD (10^{-4})↓	JSD↓
nuScenes	OpenDWM [1, 6, 42]	0.441	0.029	0.152	0.076
	LiSTAR	**0.583**	**0.017**	**0.061**	**0.056**
	Improvement	32%	41%	60%	26%
KITTI-360	OpenDWM [1, 6, 42]	0.259	0.030	0.605	0.085
	LiSTAR	**0.480**	**0.015**	**0.130**	**0.063**
	Improvement	85%	50%	79%	26%

4.2 Metrics

Evaluation Metrics. We rigorously evaluate LISTAR across three distinct tasks: point cloud reconstruction, prediction, and generation. For point cloud reconstruction, we assess both geometric accuracy and distributional fidelity. Geometric accuracy is measured by Intersection over Union (IoU) and Chamfer distance, while distributional quality is evaluated using Maximum Mean Discrepancy (MMD) and Jensen-Shannon Divergence (JSD). For point cloud prediction, our evaluation focuses on the geometric accuracy of future point clouds. We report point-wise discrepancies using Chamfer distance and a suite of L1 depth for raycasting and relative L1 error ratio, including their Mean and Median variants (L1 Med, AbsRel Med, L1 Mean, AbsRel). For LiDAR generation, we again evaluate both geometric fidelity and distributional realism. We measure Chamfer distance across multiple evaluation ranges (30 m, 40 m, and 70 m) to assess performance at varying distances. MMD and JSD quantify the overall realism and diversity of the generated sequences.

Table 2: Point cloud prediction results on nuScenes and KITTI. LiSTAR establishes a new state-of-the-art across both 1 s and 3 s prediction horizons. Compared to the baselines, our method achieves up to a 49% reduction in Chamfer distance and a 64% reduction in L1 Med.

Dataset	Method	Chamfer↓	L1 Med↓	AbsRel Med↓	L1 Mean↓	AbsRel↓
nuScenes 1 s	SPFNet [55]	2.24	-	-	4.58	34.87
	S2Net [54]	1.70	-	-	3.49	28.38
	4D-Occ [20]	1.41	0.26	4.02	1.40	10.37
	Copilot4D [64]	0.36	0.10	1.30	1.30	8.58
	LiSTAR	**0.30**	**0.05**	**0.96**	**0.76**	**4.92**
	Improvement	17%	50%	26%	42%	43%
nuScenes 3 s	SPFNet [55]	2.50	-	-	5.11	32.74
	S2Net [54]	2.06	-	-	4.78	30.15
	4D-Occ [20]	1.40	0.43	6.88	1.71	13.48
	Copilot4D [64]	0.58	0.14	1.86	1.51	10.38
	LiSTAR	**0.48**	**0.12**	**1.83**	**0.83**	**7.52**
	Improvement	17%	14%	1.6%	45%	28%
KITTI 1 s	ST3DCNN [36]	4.11	-	-	3.13	26.94
	4D-Occ [20]	0.51	0.20	2.52	1.12	9.09
	Copilot4D [64]	0.18	0.11	1.32	0.95	8.59
	LiSTAR	**0.11**	**0.04**	**0.77**	**0.62**	**3.55**
	Improvement	39%	64%	42%	35%	59%
KITTI 3 s	ST3DCNN [36]	4.19	-	-	3.25	28.58
	4D-Occ [20]	0.96	0.32	3.99	1.45	12.23
	Copilot4D [64]	0.45	0.17	2.18	1.27	11.50
	LiSTAR	**0.23**	**0.12**	**0.86**	**0.77**	**4.45**
	Improvement	49%	29%	61%	39%	61%

4.3 Reconstruction Results

As shown in Table 1, LiSTAR significantly outperforms the previous state-of-the-art method, OpenDWM, in point cloud reconstruction across both the nuScenes and KITTI-360 datasets. Our method achieves substantial gains across the board. Notably, on nuScenes, LiSTAR yields a 32% relative increase in IoU and a 60% reduction in MMD. The improvements are even more pronounced on KITTI-360, reaching an 85% increase in IoU and a 79% reduction in MMD. These results demonstrate that LiSTAR not only reconstructs scene geometry more accurately, evidenced by higher IoU and lower Chamfer Distance, but also captures the underlying data distribution with much higher fidelity, as reflected in the lower MMD and JSD scores. This comprehensive improvement validates the effectiveness of our proposed architecture for high-quality LiDAR reconstruction.

Table 3: Generation results on nuScenes and KITTI-360. LiSTAR significantly outperforms the OpenDWM baseline in both geometric accuracy and distributional fidelity. Notably, our method reduces the Chamfer distance by up to 63% and MMD by 61%. The Chamfer distance is evaluated across three spatial ranges: 30 m (magenta), 40 m (green), and 50 m (blue).

Dataset	Method	Chamfer↓	Chamfer↓	Chamfer↓	MMD (10^{-4})↓	JSD↓
nuScenes	OpenDWM [1, 6, 42]	1.59	2.51	3.49	-	-
	LiSTAR	**0.72**	**1.21**	**1.53**	**9.94**	**0.30**
	Improvement	55%	52%	56%	-	-
KITTI-360	OpenDWM [1, 6, 42]	1.88	2.57	3.35	41.1	0.31
	LiSTAR	**1.43**	**0.97**	**1.25**	**16.1**	0.31
	Improvement	24%	62%	63%	61%	0.0%

4.4 Prediction Results

The results for future point cloud prediction are presented in Table 2. Our method, LiSTAR, significantly outperforms all baselines across both the nuScenes and KITTI datasets, establishing a new state-of-the-art for both 1,s and 3,s prediction horizons. On the nuScenes dataset, LiSTAR reduces the Chamfer distance by 17% and the L1 Med by up to 50% compared to the baseline, Copilot4D. The performance gains are even more substantial on the KITTI dataset, where our method achieves up to a 49% reduction in Chamfer distance at the 3 s horizon and a 64% reduction in L1 Med at the 1 s horizon. This comprehensive improvement validates the effectiveness of our architecture in capturing complex spatio-temporal dynamics for highly accurate future predictions.

4.5 Generation Results

Table 3 presents the quantitative results for LiDAR generation on the nuScenes and KITTI-360 datasets. LiSTAR demonstrates a substantial advantage over the OpenDWM baseline across all metrics. Notably, our method achieves a 61% reduction in MMD, indicating significantly higher distributional fidelity. Furthermore, LiSTAR improves geometric accuracy by reducing the Chamfer distance by up to 63% across the evaluated spatial ranges of 30 m (magenta), 40 m (green), and 50 m (blue). These results validate the superior capability of our model in generating highly realistic LiDAR sequences.

4.6 Ablation Study

Analysis of Coordinate Representation Table 4 presents our ablation study on coordinate systems, demonstrating the clear superiority of our proposed HCS representation. HCS substantially outperforms both Cartesian and the stronger Polar coordinate baselines across all metrics. Specifically, it achieves an IoU of 0.554, marking a significant 16% relative improvement over Polar coordinates. This result strongly validates the advantage of HCS in providing a more powerful representation for LiDAR data.

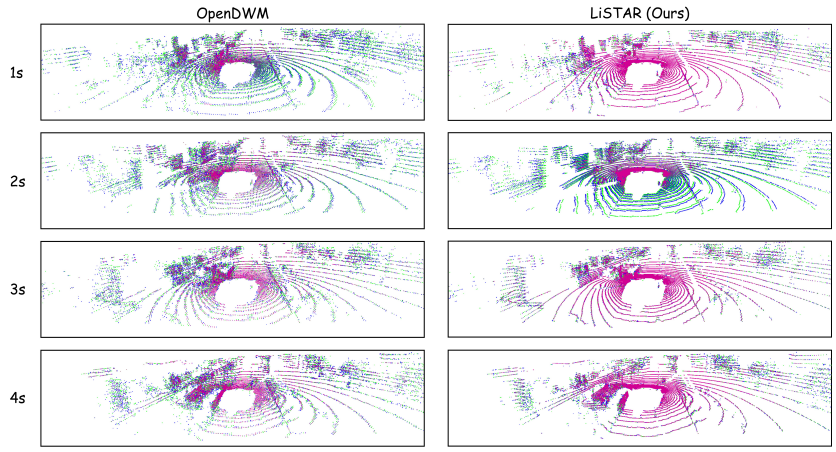

Fig. 4: Qualitative comparison of point cloud reconstruction over a 4 s horizon on the nuScenes dataset. The visualization overlays predictions with the ground truth: magenta (correct intersection), green (missed ground truth), and blue (artifacts). Our method consistently yields more complete reconstructions (denser magenta) with significantly fewer artifacts (less blue), demonstrating superior accuracy.

Table 4: Our proposed HCS coordinate achieves superior performance. The table presents a direct comparison against standard Cartesian and Polar coordinates, where our method demonstrates significant gains across all metrics. For example, it boosts IoU by 16% over the next-best polar representation. Bold denotes the best performance.

Cartesian	**Polar**	**HCS**	IoU↑	Chamfer↓	MMD (10^{-4})↓	JSD↓
✓			0.414	0.039	0.475	0.086
	✓		0.476	0.023	0.072	0.067
		✓	**0.554** (16%↑)	**0.020** (13%↓)	**0.065** (10%↓)	**0.060** (10%↓)

Effectiveness of the START Module Table 5 validates the synergistic design of our START module, demonstrating that both SRA and CSTA are critical for performance. The introduction of SRA alone provides the most significant leap, dramatically improving IoU from 0.503 to 0.554 and slashing the MMD from 0.116 to 0.065. The subsequent addition of CSTA further enhances performance across all metrics, leading to the best overall scores, including a final IoU of 0.583. This clearly shows that while SRA captures the core geometric structure, CSTA is essential for achieving the highest level of temporal and distributional fidelity.

4.7 Qualitative Results

Our qualitative results in Fig. 4 and Fig. 5 demonstrate LiSTAR's clear superiority across all tasks. In reconstruction, the baseline accumulates significant arti-

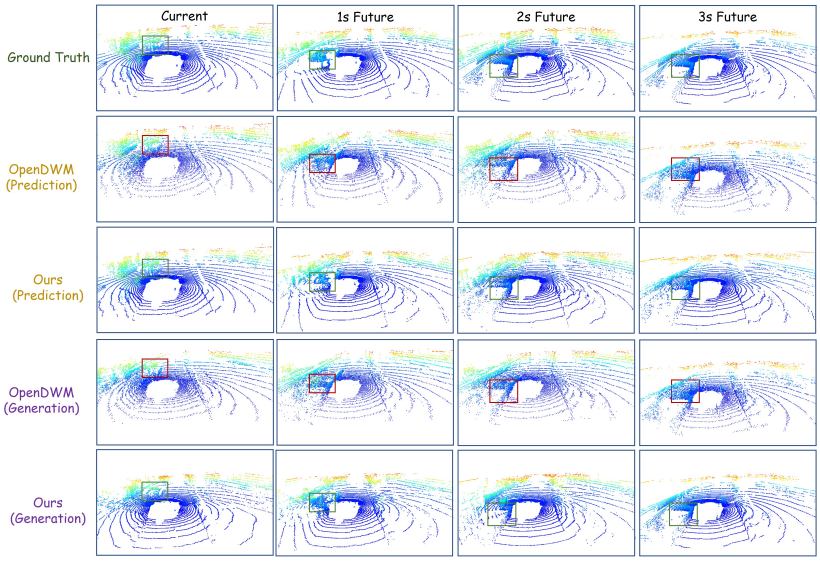

Fig. 5: Qualitative results for prediction and generation on the nuScenes dataset. We compare our method with OpenDWM against the ground truth for future horizons up to 3 s. Our method consistently produces sharper and more accurate results for both static background and dynamic objects (highlighted) compared to the baseline.

facts (blue points) over time, while our method maintains high fidelity with more true positives (magenta). Similarly, for prediction and generation, the baseline's outputs become progressively blurry and lose structural detail, whereas our results remain sharp and temporally consistent, closely matching the ground truth. This visual evidence confirms our model's advanced capability in modeling complex 4D dynamics with high fidelity. Further experimental details are provided in the Supplementary Material.

Table 5: Combining attention mechanisms in START yields superior performance. The table ablates our two attention mechanisms, showing that each provides a substantial gain over the baseline. Their combination in the START module is most effective, for instance, improving IoU from 0.503 to 0.583. Bold denotes the best performance.

SRA	CSTA	IoU↑	Chamfer↓	MMD (10^{-4})↓	JSD↓
		0.503	0.021	0.116	0.061
✓		0.554	0.020	0.065	0.060
✓	✓	**0.583**	**0.017**	**0.061**	**0.056**

5 Conclusion

In this paper, we introduced LiSTAR, a novel generative world model for high-fidelity, controllable 4D LiDAR synthesis. By unifying a novel HCS representation with a START, LiSTAR effectively preserves geometric fidelity and ensures temporal coherence. Its discrete MaskSTART framework further enables efficient, high-resolution generation conditioned on scene layouts. We have demonstrated that LiSTAR establishes a new state-of-the-art across reconstruction, prediction, and generation tasks, providing a powerful tool for creating realistic simulation environments for autonomous driving. Future work could explore multi-modal conditioning for even richer scene synthesis.

References

1. Opendwm: Open driving world models (2025), https://github.com/SenseTime-FVG/OpenDWM
2. Amini, A., Wang, T.H., Gilitschenski, I., Schwarting, W., Liu, Z., Han, S., Karaman, S., Rus, D.: VISTA 2.0: An Open, Data-driven Simulator for Multimodal Sensing and Policy Learning for Autonomous Vehicles. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2419–2426 (2022)
3. Caccia, L., Van Hoof, H., Courville, A., Pineau, J.: Deep Generative Modeling of LiDAR Data. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5034–5040 (2019)
4. 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)
5. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: MaskGIT: Masked Generative Image Transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022)
6. Chen, R., Wu, Z., Liu, Y., Guo, Y., Ni, J., Xia, H., Xia, S.: Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. arXiv preprint arXiv:2412.04842 (2024)
7. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In: European conference on computer vision. pp. 628–644 (2016)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
9. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An Open Urban Driving Simulator. In: Conference on robot learning. pp. 1–16 (2017)
10. Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
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. arXiv preprint arXiv:2310.02601 (2023)

12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
13. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean Flows for One-step Generative Modeling. arXiv preprint arXiv:2505.13447 (2025)
14. Graham, B., Engelcke, M., Van Der Maaten, L.: 3D Semantic Segmentation With Submanifold Sparse Convolutional Networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9224–9232 (2018)
15. 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)
16. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. 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)
18. Hu, Q., Zhang, Z., Hu, W.: RangeLDM: Fast Realistic LiDAR Point Cloud Generation. In: European Conference on Computer Vision. pp. 115–135 (2024)
19. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: VBench: Comprehensive Benchmark Suite for Video Generative Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
20. Khurana, T., Hu, P., Held, D., Ramanan, D.: Point Cloud Forecasting as a Proxy for 4D Occupancy Forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1116–1124 (2023)
21. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114 (2013)
22. Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robo3D: Towards Robust and Reliable 3D Perception against Corruptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19994–20006 (2023)
23. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: PointPillars: Fast Encoders for Object Detection From Point Clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
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, H., Hou, Y., Xing, X., Ma, Y., Sun, X., Zhang, Y.: OccMamba: Semantic Occupancy Prediction with State Space Models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11949–11959 (2025)
26. Li, J., Liu, J., Huang, Q.: PointDMM: A Deep-Learning-Based Semantic Segmentation Method for Point Clouds in Complex Forest Environments. Forests **14**(12), 2276 (2023)
27. Liang, A., Liu, Y., Yang, Y., Lu, D., Li, L., Kong, L., Zhao, H., Ooi, W.T.: LiDARCrafter: Dynamic 4D World Modeling from LiDAR Sequences. arXiv preprint arXiv:2508.03692 (2025)
28. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3292–3310 (2022)

29. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A Structured Self-attentive Sentence Embedding. arXiv preprint arXiv:1703.03130 (2017)
30. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow Matching for Generative Modeling. arXiv preprint arXiv:2210.02747 (2022)
31. Liu, Y., Kong, L., Cen, J., Chen, R., Zhang, W., Pan, L., Chen, K., Liu, Z.: Segment Any Point Cloud Sequences by Distilling Vision Foundation Models. Advances in Neural Information Processing Systems **36**, 37193–37229 (2023)
32. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
33. Liu, Z., Tang, H., Zhao, S., Shao, K., Han, S.: PVNAS: 3D Neural Architecture Search With Point-Voxel Convolution. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8552–8568 (2021)
34. Manivasagam, S., Wang, S., Wong, K., Zeng, W., Sazanovich, M., Tan, S., Yang, B., Ma, W.C., Urtasun, R.: LiDARsim: Realistic LiDAR Simulation by Leveraging the Real World. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11167–11176 (2020)
35. Mei, J., Hu, T., Yang, X., Wen, L., Yang, Y., Wei, T., Ma, Y., Dou, M., Shi, B., Liu, Y.: DreamForge: Motion-Aware Autoregressive Video Generation for Multi-View Driving Scenes. arXiv preprint arXiv:2409.04003 (2024)
36. Mersch, B., Chen, X., Behley, J., Stachniss, C.: Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In: Conference on Robot Learning. pp. 1444–1454. PMLR (2022)
37. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. Communications of the ACM **65**(1), 99–106 (2021)
38. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 4213–4220 (2019)
39. Nakashima, K., Kurazume, R.: Learning to Drop Points for LiDAR Scan Synthesis. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 222–229 (2021)
40. Nakashima, K., Kurazume, R.: LiDAR Data Synthesis with Denoising Diffusion Probabilistic Models. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 14724–14731 (2024)
41. Nakashima, K., Liu, X., Miyawaki, T., Iwashita, Y., Kurazume, R.: Fast LiDAR Data Generation with Rectified Flows. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 10057–10063. IEEE (2025)
42. Ni, J., Guo, Y., Liu, Y., Chen, R., Lu, L., Wu, Z.: Maskgwm: A generalizable driving world model with video mask reconstruction. arXiv preprint arXiv:2502.11663 (2025)
43. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
44. Qi, C.R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L.J.: Volumetric and Multi-View CNNs for Object Classification on 3D Data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5648–5656 (2016)

45. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. Advances in neural information processing systems **30** (2017)
46. Ran, H., Guizilini, V., Wang, Y.: Towards Realistic Scene Generation with LiDAR Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14738–14748 (2024)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
48. 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)
49. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. arXiv preprint arXiv:2010.02502 (2020)
50. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. arXiv preprint arXiv:2011.13456 (2020)
51. Van Den Oord, A., Vinyals, O., et al.: Neural Discrete Representation Learning. Advances in neural information processing systems **30** (2017)
52. Wang, L., Zheng, W., Ren, Y., Jiang, H., Cui, Z., Yu, H., Lu, J.: OccSora: 4D Occupancy Generation Models as World Simulators for Autonomous Driving. arXiv preprint arXiv:2405.20337 (2024)
53. 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)
54. Weng, X., Nan, J., Lee, K.H., McAllister, R., Gaidon, A., Rhinehart, N., Kitani, K.M.: S2Net: Stochastic Sequential Pointcloud Forecasting. In: European Conference on Computer Vision. pp. 549–564. Springer (2022)
55. Weng, X., Wang, J., Levine, S., Kitani, K., Rhinehart, N.: Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. In: Conference on robot learning. pp. 11–20. PMLR (2021)
56. Wu, W., Guo, X., Tang, W., Huang, T., Wang, C., Chen, D., Ding, C.: DriveScape: Towards High-Resolution Controllable Multi-View Driving Video Generation. arXiv preprint arXiv:2409.05463 (2024)
57. Xiong, Y., Ma, W.C., Wang, J., Urtasun, R.: Learning Compact Representations for LiDAR Completion and Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1074–1083 (2023)
58. Xu, X., Kong, L., Shuai, H., Zhang, W., Pan, L., Chen, K., Liu, Z., Liu, Q.: 4D Contrastive Superflows are Dense 3D Representation Learners. In: European Conference on Computer Vision. pp. 58–80. Springer (2024)
59. Yang, B., Luo, W., Urtasun, R.: PIXOR: Real-Time 3D Object Detection From Point Clouds. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7652–7660 (2018)
60. Yang, Z., Chen, L., Sun, Y., Li, H.: Visual Point Cloud Forecasting enables Scalable Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14673–14684 (2024)
61. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: MAGVIT: Masked Generative Video Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10459–10469 (2023)

62. Yu, S., Sohn, K., Kim, S., Shin, J.: Video Probabilistic Diffusion Models in Projected Latent Space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18456–18466 (2023)
63. Zhang, J., Zhang, F., Kuang, S., Zhang, L.: NeRF-LiDAR: Generating Realistic LiDAR Point Clouds with Neural Radiance Fields. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7178–7186 (2024)
64. Zhang, L., Xiong, Y., Yang, Z., Casas, S., Hu, R., Urtasun, R.: Copilot4D: Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion. arXiv preprint arXiv:2311.01017 (2023)
65. 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)
66. 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: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10412–10420 (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 (2024)
68. Zuo, S., Zheng, W., Huang, Y., Zhou, J., Lu, J.: GaussianWorld: Gaussian World Model for Streaming 3D Occupancy Prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6772–6781 (2025)
69. Zyrianov, V., Che, H., Liu, Z., Wang, S.: LidarDM: Generative LiDAR Simulation in a Generated World. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6055–6062 (2025)
70. Zyrianov, V., Zhu, X., Wang, S.: Learning to Generate Realistic LiDAR Point Clouds. In: European Conference on Computer Vision. pp. 17–35 (2022)