3D and 4D World Modeling: A Survey

Lingdong Kong*, Wesley Yang*, Jianbiao Mei*, Youquan Liu*, Ao Liang*, Dekai Zhu*, Dongyue Lu*, Wei Yin*, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, Ziwei Liu

Abstract—World modeling has become a cornerstone in AI research, enabling agents to understand, represent, and predict the dynamic environments they inhabit. While prior work largely emphasizes generative methods for 2D image and video data, they overlook the rapidly growing body of work that leverages **native 3D and 4D representations** such as RGB-D imagery, occupancy grids, and LiDAR point clouds for large-scale scene modeling. At the same time, the absence of a standardized definition and taxonomy for "world models" has led to fragmented and sometimes inconsistent claims in the literature. This survey addresses these gaps by presenting the first comprehensive review explicitly dedicated to 3D and 4D world modeling and generation. We establish precise definitions, introduce a structured taxonomy spanning video-based (**VideoGen**), occupancy-based (**OccGen**), and LiDAR-based (**LiDARGen**) approaches, and systematically summarize datasets and evaluation metrics tailored to 3D/4D settings. We further discuss practical applications, identify open challenges, and highlight promising research directions, aiming to provide a coherent and foundational reference for advancing the field. A systematic summary of existing literature is available at https://github.com/worldbench/survey.

Index Terms—World Models; Video Generation; 3D Generation; 4D Generation; 3D Scene Understanding; Spatial Intelligence

1 Introduction

World modeling has emerged as a fundamental task in AI and robotics, aiming towards the ability to understand, represent, and anticipate the dynamic environments they inhabit [1], [2], [3]. Recent advancements in generative modeling techniques, including VAEs, GANs, diffusion models, and autoregressive models, have significantly enriched the field by enabling sophisticated generation and prediction capabilities [4], [5].

Much of this progress, however, has been centered on 2D data, primarily images or videos [6], [7], [8]. Real-world scenarios, in contrast, are inherently in 3D space and dynamic, often requiring models that leverage **native 3D and 4D representations**. These include RGB-D imagery [9], [10], [11], occupancy grids [12], [13], [14], and LiDAR point clouds [15], [16], [17], as well as their sequential forms that capture temporal dynamics [18], [19]. These modalities offer explicit geometry and physical grounding, which are indispensable for embodied and safety-critical systems such as autonomous driving and robotics [20], [21], [22], [23], [24], [25], [26].

Beyond these native formats, world modeling has also been explored in adjacent domains [27], [28], [29]. Some works address video, panoramic, or mesh-based data, with systems of this kind providing large-scale, general-purpose video-mesh generation capabilities [30], [31]. In parallel, another line of research focuses on 3D object generation for asset creation, which specializes in controllable and high-fidelity object synthesis [32], [33], [34]. Meanwhile, industrial projects from leading companies have launched ambitious world modeling initiatives that target practical applications ranging from interactive robotics and immersive simulation to large-scale digital twins [35], [36], [37], [38], [39], [40], underscoring the growing importance of this field in both academia and industry.

Despite this momentum, the term "world model" itself remains ambiguous, with inconsistent usage across the literature [27], [41], [42]. Some works narrowly interpret it as generative models for sensory data (*e.g.*, images and videos), while others broaden the scope to include predictive forecasting, simulators, and decision-making frameworks [43], [44], [45], [46], [47]. Moreover, existing surveys largely emphasize 2D or vision-only modalities [6], [48], leaving the unique challenges and opportunities of native 3D and 4D data underexplored. This has led to a fragmented body of literature lacking a unified framework or taxonomy.

- **Why native 3D and 4D matters?** Unlike 2D projections, native 3D/4D signals encode metric geometry, visibility, and motion in the coordinates where physics acts [18], [49]. This makes them *first-class carriers* of constraints needed for actionable modeling: multi-view and egocentric consistency, rigid-body and non-rigid kinematics, scene-scale occlusion reasoning, and map/topology adherence. In safety-critical settings, agents must not only produce photorealistic frames but also obey geometry, causality, and controllability; RGB-D, occupancy, and LiDAR provide the inductive bias to satisfy these requirements. Sec. 2 will formalize these representations and the conditioning signals ($\mathcal{C}_{\text{geo}}, \mathcal{C}_{\text{act}}, \mathcal{C}_{\text{sem}}$) we use throughout the survey.
- **Position in the broader landscape.** The adjacent lines – video/panorama/mesh world models [30], [31] and object-

- *L. Kong, A. Liang, D. Lu, L. Li, and W. T. Ooi are with the National University of Singapore. L. Kong is also with CNRS@CREATE, Singapore.*
- *J. Mei, S. Wang, Y. Liu, and J. Zhu are with Zhejiang University.*
- *W. Yin is with Horizon Robotics.*
- *D. Zhu is with the Technical University of Munich.*
- *X. Hu, M. Jia, J. Deng, and S. Gao are with HKUST.*
- *K. Zhang is with Tsinghua University.*
- *Y. Wu is with Nanjing University of Science and Technology.*
- *T. Yan is with the University of Macau.*
- *W. Yang and L. Pan are with Shanghai AI Laboratory.*
- *S. C. H. Hoi is with Alibaba Group and Singapore Management University.*
- *Z. Liu is with Nanyang Technological University, Singapore.*
- *The corresponding authors are Y. Liu, J. Zhu, W. T. Ooi, and Z. Liu.*
- *(∗) These authors contributed equally to this work.*

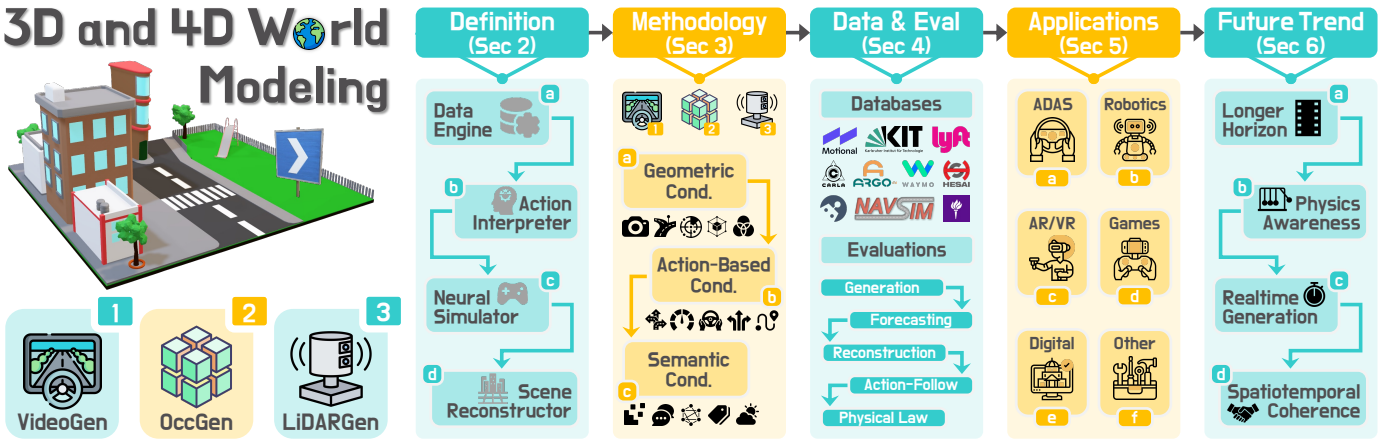

Fig. 1: **Outline of the survey.** This work focuses on **native 3D and 4D representations**: video streams, occupancy grids, and LiDAR point clouds, guided by geometric ($\mathcal{C}_{\text{geo}}$), action-based ($\mathcal{C}_{\text{act}}$), and semantic ($\mathcal{C}_{\text{sem}}$) conditions (Sec. 2). Methods are framed under two paradigms, *generative* (synthesis from observations and conditions) and *predictive* (forecasting from history and actions), and grouped into four functional types (Sec. 3). We cover three modality tracks and standardize evaluations (Sec. 4), applications (Sec. 5), and future trends (Sec. 6) across diverse generation, forecasting, and downstream tasks.

centric 3D asset generators [32], [33], [34] – are complementary: they supply appearance, topology, and assets, while native 3D/4D world models supply geometry-grounded dynamics and interaction [19], [50]. Practical systems increasingly compose these capabilities: mesh/panorama worlds initialized from assets, then *driven* by occupancy- or LiDAR-based dynamics, or video models constrained by 3D priors for view and motion correctness. Our scope centers on the latter – native 3D/4D – while acknowledging and cross-referencing where cross-pollination occurs.

- **From conditions to functions.** A common pain point in the field is conflating *"what the model consumes"* (conditions) with *"what the model does"* (function). We therefore separate the roles of geometry/action/semantics conditions (Table 1) from functional types. Sec. 3 organizes methods by representation modality –**VideoGen**, **OccGen**, **LiDARGen** – and then by **four** functional roles: [1]*Data Engines* (diverse scene synthesis under $\mathcal{C}_{\text{geo}}, \mathcal{C}_{\text{sem}}, \mathcal{C}_{\text{act}}$), [2]*Action Interpreters* (forecasting under $\mathcal{C}_{\text{act}}$ with history), [3]*Neural Simulators* (closed-loop rollouts with policy-in-the-loop), and [4]*Scene Reconstructors* (completion/retargeting from partial 3D/4D observations). This decoupling lets us compare heterogeneous methods on common axes of fidelity, consistency, controllability, and scalability.

Contributions. To address the aforementioned gaps, this survey presents the first comprehensive review specifically dedicated to **3D and 4D world modeling and generation**. The primary contributions of this survey are **threefold**:

- We establish precise definitions for "world models" and "3D/4D world modeling", providing the research community with consistent terminology and conceptual clarity.
- We propose a hierarchical taxonomy of methodologies, categorizing current approaches based on their representation modalities – namely, world modeling based on VideoGen, OccGen, and LiDARGen models.
- We provide extensive coverage of datasets and evaluation protocols specifically tailored for 3D and 4D scenarios, enabling a thorough benchmarking of existing and future world modeling and generation approaches.

Scope. Distinct from previous surveys, which predominantly focus on 2D generative models [48], [51], [52] or broadly define world modeling within limited contexts [53], [54], [55], [56], [57], this survey explicitly targets methodologies that utilize native 3D and 4D representations. This specialized focus includes approaches leveraging RGB-D, volumetric occupancy grids, LiDAR point clouds, and their spatiotemporal forms. By highlighting these modalities, our survey not only fills a critical knowledge gap but also serves as a foundational reference for researchers aiming to develop robust and generalizable 3D/4D generative models.

Organization. The remainder of this survey is organized as follows. Sec. 2 provides preliminaries, detailing fundamental concepts, definitions, and key generative paradigms relevant to world modeling. Sec. 3 introduces a new and hierarchical taxonomy, detailing VideoGen, OccGen, and LiDARGen methodologies, providing comparative analyses and insights into their respective strengths and limitations. Sec. 4 systematically summarizes and categorizes widely used datasets and evaluation metrics critical for world modeling tasks, as well as benchmarking recent methods in this related area. Sec. 5 reviews practical applications of 3D and 4D world models across autonomous driving, robotics, and simulation environments. Sec. 6 discusses major challenges and highlights promising future research directions, paving the way for continued innovation in the field. Finally, Sec. 7 concludes the key discussions drawn in this survey.

2 Preliminaries

In this section, we define critical concepts and establish unified mathematical notations essential for understanding 3D and 4D world modeling. This includes detailed descriptions of the key representations, definitions of generative and predictive world models, and model categorizations.

2.1 3D and 4D Representations

To systematically analyze 3D/4D world models, we first introduce the fundamental scene representations that serve

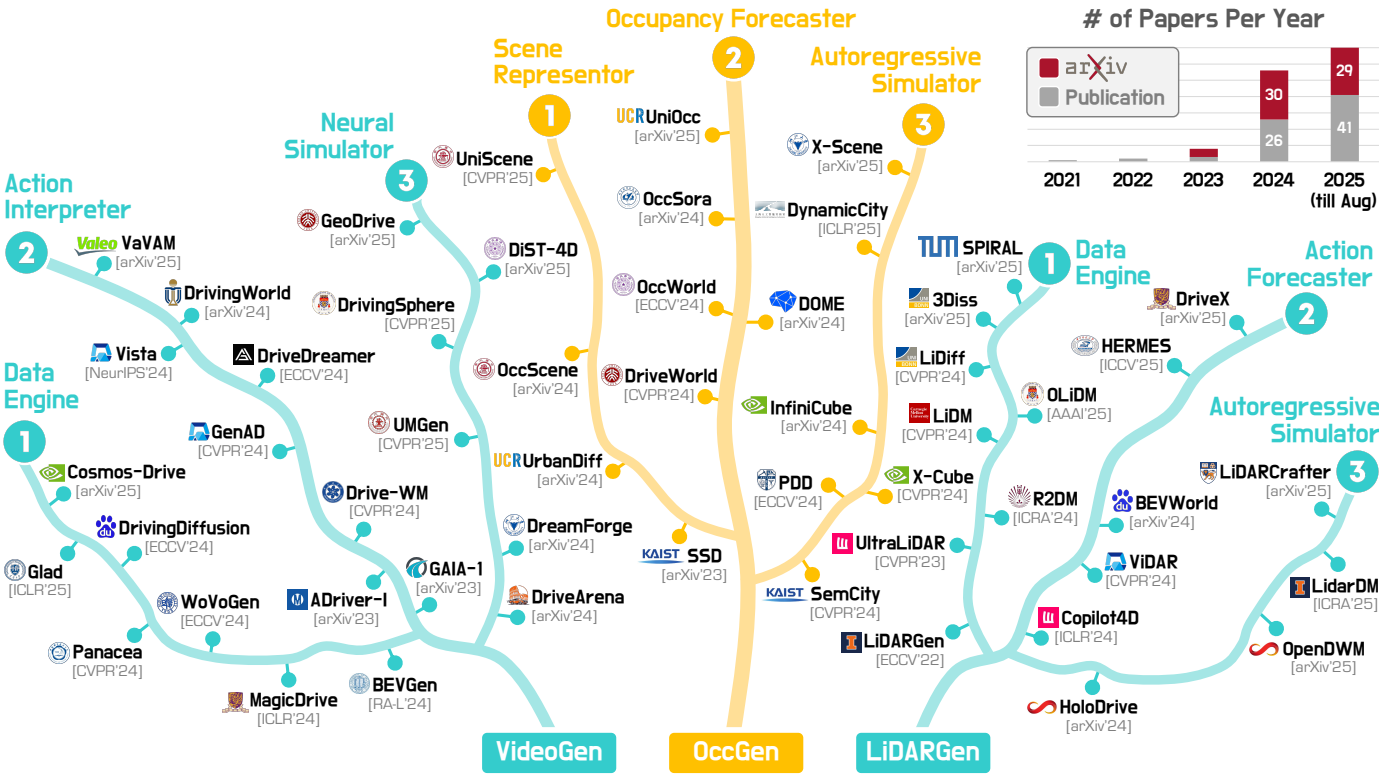

Fig. 2: Summary of representative video-based generation **(VideoGen)**, occupancy-based generation **(OccGen)**, and LiDAR-based generation **(LiDARGen)** models from existing literature. For the complete list of related methods and discussions on their specifications, configurations, and technical details, kindly refer to Sec. 3.1, Sec. 3.2, and Sec. 3.3, respectively.

as inputs, outputs, or intermediate states in generation and prediction. These representations differ in how they capture spatial geometry, temporal dynamics, and semantic context.

Video Streams. A video is denoted as $\mathbf{x}_v \in \mathbb{R}^{T\times H\times W\times C}$, where T is the number of frames, and H, W, C are the frame height, width, and channels. Unlike conventional 2D videos, 3D/4D modeling emphasizes *geometric coherence* and *temporal consistency* to ensure physically plausible simulations and accurate forecasting [58], [59], [60].

Occupancy Grids. A static occupancy grid is represented as $\mathbf{x}_o \in \{0,1\}^{X\times Y\times Z}$, where each voxel indicates whether a location is occupied [12], [61]. Sequential occupancy grids $\mathbf{x}_o^t \in \{0,1\}^{T\times X\times Y\times Z}$ extend this into 4D, capturing scene evolution over time. Such voxelized geometry enforces spatial constraints, making them well-suited for physics-consistent scene generation.

LiDAR Point Clouds. A LiDAR-acquired scan is expressed as $\mathbf{x}_l = \{(x_i, y_i, z_i)\}_{i=1}^{N}$, where (x_i, y_i, z_i) are the Cartesian coordinates in 3D space [62]. Sequential LiDAR $\mathbf{x}_l^t = \{(x_i, y_i, z_i, t_i)\}_{i=1}^{N_t}$ further records the timestamp t_i, enabling precise modeling of motion and interactions [63], [64]. Unlike RGB images, LiDAR captures geometry directly and remains robust to texture, lighting, or weather variations [24], [65].

Neural Representations. Implicit scene encodings, such as neural radiance fields (NeRF) and Gaussian splatting (GS), model continuous volumetric fields or explicit Gaussian primitives. NeRF maps a ray origin $\mathbf{r}$ and direction $\mathbf{d}$ to color $\mathbf{c}$ and density σ, while GS represents the scene as a set of Gaussians parameterized by position, covariance, and color. Temporal extensions add dynamic components, enabling realistic 4D reconstructions and simulations.

2.2 Definition of World Modeling in 3D and 4D

The above scene representations form the structural backbone of 3D/4D world models. In practice, generating or forecasting them requires additional *conditions* – auxiliary signals that constrain spatial structure, describe agent behavior, or define high-level semantics. As summarized in Table 1, these conditions are typically grouped into:

- **geometric** $\mathcal{C}_{\text{geo}}$**:** specifying spatial layout such as camera pose, depth maps, or occupancy volumes;
- **action-based** $\mathcal{C}_{\text{act}}$**:** describing ego-vehicle or agent motion via trajectories, control commands, or navigation goals;
- **semantic** $\mathcal{C}_{\text{sem}}$**:** providing abstract scene intent such as textual prompts, scene graphs, or environment attributes.

These signals can be used independently or in combination, shaping the realism, controllability, and diversity of the generated or forecasted scenes in 3D and 4D.

2.2.1 Model Definitions

Depending on the modeling objective, 3D/4D world models generally fall into two complementary paradigms:

Generative World Models focus on synthesizing plausible scenes from scratch or from partial observations, guided by multimodal conditions. This process can be formulated as:

$$\mathcal{G}(\mathbf{x}_i, \mathcal{C}_{\text{geo}}, \mathcal{C}_{\text{act}}, \mathcal{C}_{\text{sem}}) \rightarrow \mathcal{S}_g, \tag{1}$$

where $\mathbf{x}_i$ denotes the optional input representation, with $i \in \{\varnothing, v, o, l\}$, *e.g.*, noise, partial video, occupancy, or LiDAR

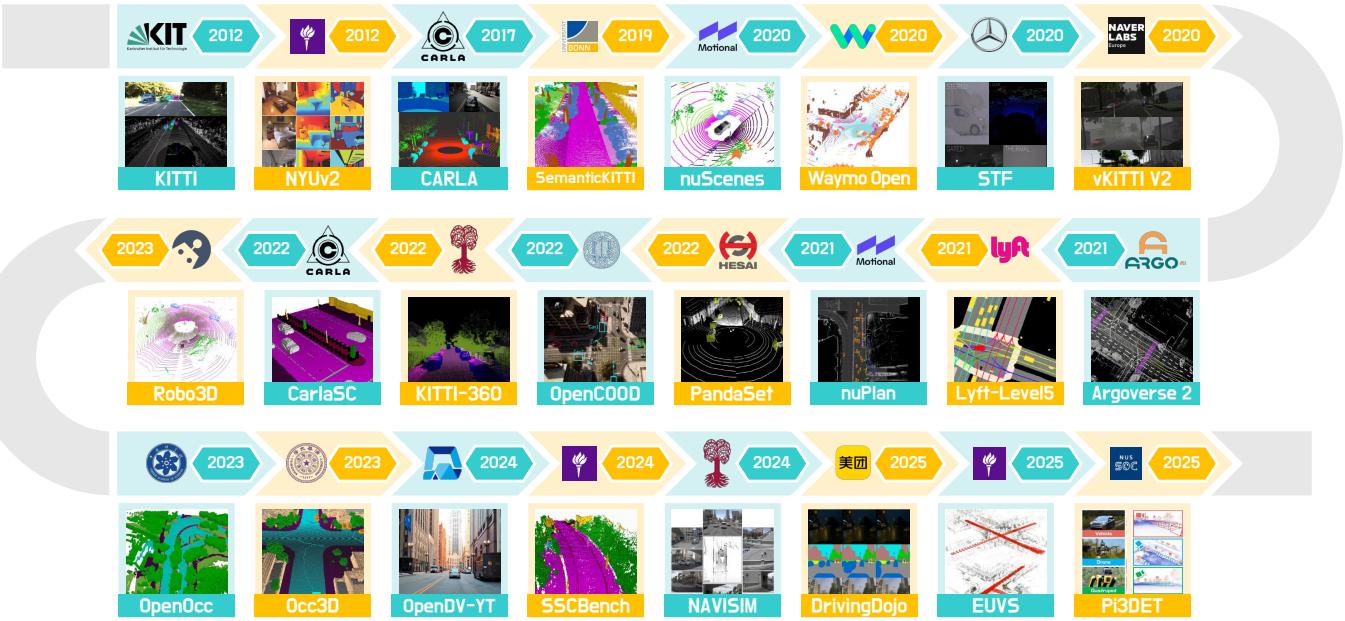

Fig. 3: Summary of existing **datasets and benchmarks** used for training and evaluating **VideoGen**, **OccGen**, and **LiDARGen** models. For detailed dataset configurations and statistics, kindly refer to Table 5. Images adopted from the original papers.

data. $\mathcal{C}_{\text{geo}}$, $\mathcal{C}_{\text{act}}$, and $\mathcal{C}_{\text{sem}}$ correspond to the geometric, action, and semantic conditions. The output $\mathcal{S}_g$ is a generated 3D/4D scene, such as a video sequence, occupancy grid, or LiDAR sweep sequence.

Predictive World Models instead aim to forecast the future evolution of the scene based on historical observations, often under action conditions that describe planned or executed agent behavior. This process can be formulated as:

$$\mathcal{P}(\mathbf{x}_i^{-t:0}, \mathcal{C}_{\text{act}}) \rightarrow \mathcal{S}_p^{1:k}, \tag{2}$$

where $\mathbf{x}_i^{-t:0}$ represents observations from the past t steps to the current step, and $\mathcal{C}_{\text{act}}$ encodes agent actions (*e.g.*, control commands or planned trajectories). The model outputs $\mathcal{S}_p^{1:k}$, the predicted scene representations over k future steps.

Together, these two paradigms capture the dual capability of world models: the ability to *imagine* diverse and controllable worlds (generative), and to *anticipate* their plausible future evolution under specific conditions (predictive).

2.2.2 Model Categorizations

Building on the generative and predictive paradigms, existing approaches can be further divided into **four functional types**. They differ in how they utilize historical observations, the nature of conditioning signals ($\mathcal{C}_{\text{geo}}, \mathcal{C}_{\text{act}}, \mathcal{C}_{\text{sem}}$), and whether they operate in an open-loop or closed-loop setting.

Type 1: Data Engines

Generate diverse 3D/4D scenes from geometric and semantic cues, optionally with action conditions.

- **Inputs:** $\mathcal{C}_{\text{geo}}$ (geometric cond.), $\mathcal{C}_{\text{act}}$ (action cond., optional), and $\mathcal{C}_{\text{sem}}$ (semantic cond.)
- **Output:** $\mathcal{S}_g$ (generated scene)

Focus on *plausibility* and *diversity* for large-scale data augmentation and scenario creation.

Type 2: Action Interpreters

Forecast future 3D/4D world states from historical observations under given action conditions.

- **Inputs:** $\mathbf{x}_i^{-t:0}$ (historical observations) and $\mathcal{C}_{\text{act}}$ (action cond.)
- **Output:** $\mathcal{S}_p^{1:k}$ (predicted sequence)

Enable *action-aware forecasting* for trajectory planning, behavior prediction, and policy evaluation.

Type 3: Neural Simulators

Iteratively simulate closed-loop agent-environment interactions by generating successive scene states.

- **Inputs:** $\mathcal{S}_g^t$ (current scene state) and π_{agent} (agent policy)
- **Output:** $\mathcal{S}_g^{t+1}$ (next scene state)

Support *interactive simulation* for autonomous driving, robotics, and immersive XR training.

Type 4: Scene Reconstructors

Recover complete and coherent 3D/4D scenes from partial, sparse, or corrupted observations.

- **Inputs:** $\mathbf{x}_i^p$ (partial observations) and $\mathcal{C}_{\text{geo}}$ (optional geometric cond.)
- **Output:** $\hat{\mathcal{S}}_g$ (completed scene)

Facilitate *interactive tasks* on high-fidelity mapping, digital twin restoration, and post-event analysis.

Together, these four categories outline the functional landscape of 3D/4D world modeling. While all aim to produce physically and semantically coherent scenes, they differ in how they leverage past observations, conditioning

TABLE 1: Summary of the rich collection of conditions used by existing **VideoGen**, **OccGen**, and **LiDARGen** models. The conditions are categorized into **three** main groups: geometric conditions, action-based conditions, and semantic conditions. The tasks are ❶ video generation (Sec. 3.1), ❷ occupancy generation (Sec. 3.2), and ❸ LiDAR generation (Sec. 3.3).

Group	Condition	Definition	Task
Geometry ($\mathcal{C}_{\text{geo}}$)	C Camera Pose	Position and orientation of the camera in world coordinates, controlling viewpoint	❶
	D Depth Map	Per-pixel depth values providing scene geometry constraints	❶
	B BEV Map	Bird's-eye-view geometric representation of the scene	❶ ❷ ❸
	H HD Map	High-resolution semantic map with detailed road layout and traffic elements	❶ ❸
	3 3D Bounding Box	Object bounding boxes in 3D, defining positions, sizes, and orientations of objects	❶ ❸
	F Flow Field	Optical or scene flow encoding per-pixel or per-point motion between frames	❶ ❷
	P Past Occupancy	Historical occupancy grids or voxel maps capturing prior scene geometry	❷ ❸
	L LiDAR Pattern	Sensor scan configuration including beam count, FOV, and resolution	❸
	O Object Coordinate	Set of Cartesian coordinates of instances from LiDAR point clouds	❶ ❸
	P Partial Point Cloud	Incomplete LiDAR point set capturing only a subset of the full 3D scene geometry	❸
	R RGB Frame	Single color image frame from a monocular or multi-camera setup	❶ ❷ ❸
	S Surface Mesh	Triangular mesh or equivalent explicit geometry representation of the scene	❶ ❷ ❸
Action ($\mathcal{C}_{\text{act}}$)	T Ego-Trajectory	The planned or recorded path of the ego vehicle over time	❶ ❷ ❸
	V Ego-Velocity	The speed and direction of the ego movement	❶ ❷
	A Ego-Acceleration	Rate of change of ego velocity, describing linear acceleration or deceleration	❶
	S Ego-Steering	The steering angle or input controlling the ego direction	❶
	C Ego-Command	The control instructions given to the ego vehicle	❶ ❷ ❸
	R Route Plan	High-level navigation path through the environment, often from a planner	❶
	A Action Token	Encoded discrete actions or instructions influencing scene evolution	❶
	S Scan Path	Predefined movement or sweep pattern during LiDAR acquisition	❸
Semantics ($\mathcal{C}_{\text{sem}}$)	S Semantic Mask	Pixel-/occupancy-/point-wise semantic categories	❶ ❷ ❸
	T Text Prompt	Natural language input specifying scene attributes, objects, or actions	❶ ❷ ❸
	G Scene Graph	Graph representation of scene entities and their spatial/semantic relationships	❷ ❸
	O Object Label	Class category annotation assigned to an object instance in the scene	❶ ❷ ❸
	W Weather Tag	Discrete label describing environmental conditions such as sunny, rainy, or foggy	❶ ❷
	M Material Tag	Classification of surface materials influencing appearance or LiDAR reflectance	❶ ❸

signals, and interaction loops – serving applications ranging from large-scale data synthesis and policy evaluation to interactive simulation and scene restoration.

2.3 Generative Models

Generative models form the algorithmic core of 3D/4D world modeling, enabling agents to *learn*, *imagine*, and *forecast* sensory data under diverse conditions. They provide the mechanisms to synthesize realistic and physically plausible scenes, with different paradigms offering distinct trade-offs in quality, controllability, and efficiency. Representative families include variational autoencoders, generative adversarial networks, diffusion models, and autoregressive models.

Variational Autoencoders (VAEs) [66] learn a structured latent space via probabilistic encoding and decoding. Given input $\mathbf{x}$, the encoder defines a variational posterior $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_\phi(\mathbf{x}), \text{diag}(\sigma^2_\phi(\mathbf{x})))$ and samples $\mathbf{z}$ using the reparameterization trick: $\mathbf{z} = \mu_\phi(\mathbf{x}) + \sigma_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. The decoder $p_\theta(\mathbf{x}|\mathbf{z})$ reconstructs the input, and the model is trained to maximize the variational lower bound that balances reconstruction fidelity and latent regularization:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})). \tag{3}$$

VAEs offer stable training and interpretable latent spaces, but may produce blurrier samples compared to other paradigms.

Generative Adversarial Networks (GANs) [67] generate data via a min–max game between a generator G_θ and discriminator D_ϕ. The generator maps latent variables $\mathbf{z} \sim p(\mathbf{z})$ to the data space, aiming to fool D_ϕ, while the discriminator distinguishes real from synthetic samples:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]. \tag{4}$$

GANs can produce high-fidelity result samples but often suffer from training instability and mode collapse issues.

Diffusion Models (DMs) [68], [69] learn to reverse a gradual noising process. The forward process corrupts $\mathbf{x}_0$ into $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ via $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$, where β_t follows a variance schedule. The reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is trained to denoise, minimizing:

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon},t}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2]. \tag{5}$$

DMs provide strong stability and sample quality, though inference can be slow due to iterative sampling.

Autoregressive Models (ARs) [70], [71] factorize the joint distribution as $p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid x_{<i})$, predicting each element conditioned on all previous ones. Transformer-based ARs offer exact likelihood estimation and flexible sequence modeling, but suffer from slow generation since samples are produced sequentially. Recent advances have adapted ARs to spatial and temporal tokens, making them well-suited for structured 3D scene generation and forecasting.

Summary. These paradigms form the algorithmic backbone for world models. Their differences in structure, training stability, and inference efficiency directly shape how 3D environments can be synthesized, forecasted, and controlled. As we move into native 3D/4D domains, these trade-offs are magnified, since scalability, controllability, and multi-modal integration are critical to constructing reliable world models for embodied AI and simulation.

3 Methods: A Hierarchical Taxonomy

In this section, we standardize and categorize existing 3D and 4D world modeling approaches based on their representation modalities. This includes descriptions and discussions of world modeling based on **Video Generation** (Sec. 3.1), **Occupancy Generation** (Sec. 3.2), and **LiDAR Generation** (Sec. 3.3) models, respectively.

3.1 World Modeling from Video Generation

Video-based generation has emerged as a new paradigm, offering visual cues and temporal dynamics to model complex real-world scenarios. By generating multi-view or egocentric video sequences, these models can synthesize training data, predict future outcomes, and create interactive simulation environments. Based on their primary function, existing methods can be grouped into **three** categories: ❶ **Data Engines**, ❷ **Action Interpreters**, and ❸ **Neural Simulators**. Table 2 summarizes existing models under these domains.

3.1.1 Data Engines

Generative 3D data engines focus on generating diverse and controllable driving scenes to support perception, planning, and simulation [20], [72], [73], [74], [75], [76], [77], [78]. Research in this direction covers three major applications.

Perception Data Augmentation. Generative scene synthesis alleviates real-world data scarcity and addresses long-tail perception challenges. Early work focused on BEV-guided realistic street scenes. BEVGen [72] uses an autoregressive transformer and cross-view transformation to produce spatially consistent surrounding images aligned with a given BEV layout. BEVControl [73] centers on diffusion models to boost the quality of synthetic data, particularly for augmenting challenging long-tail scenarios. Subsequently, MagicDrive [20] made significant progress in driving scene generation and data augmentation, combining 3D geometry and semantic descriptions, and camera poses to generate high-fidelity images. Later work introduced finer conditioning. For instance, SyntheOcc [75] uses 3D semantic multi-plane images for comprehensive, spatially aligned conditioning, and PerLDiff [79] proposes perspective-layout diffusion models that fully leverage perspective 3D geometry to enhance realism and consistency. On the other hand, approaches such as Panacea [80], DrivingDiffusion [81], and SubjectDrive [82] introduce 4D attention, keyframes, and subject control to improve the temporal consistency and data diversity of 3D controllable multi-view videos. NoiseController [78] proposes multi-level noise decomposition and multi-frame collaborative denoising to enhance spatiotemporal coherence. For long-horizon video generation, DiVE [76], MagicDrive-V2 [74], and Cosmos-Drive [35] leverage the flexibility and scalability of DiT to produce longer videos. Glad [83] uses latent-variable propagation, and STAGE [84] uses hierarchical temporal feature transfer to generate long videos in a streaming fashion. Others like UniScene [77] and BEVWorld [85] explore multi-modal data synthesis to broaden applications, supporting downstream perception tasks that leverage information from multiple modalities. These advances enable robust, scalable autonomous driving perception systems by delivering diverse, controllable, and long-horizon training data that capture real-world variability.

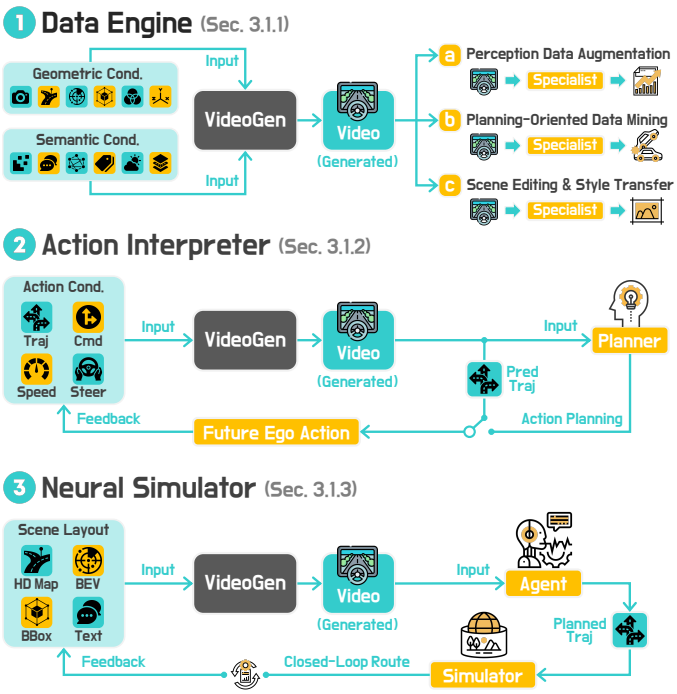

Fig. 4: The categorization of **VideoGen** models based on functionalities, including data engines (Sec. 3.1.1), action interpreters (Sec. 3.1.2), and neural simulators (Sec. 3.1.3).

Planning-Oriented Data Mining. Beyond perception, data engines also mine rare and safety-critical scenarios for planning. Delphi [86] employed a diffusion-based long video generation framework and a failure-case-driven approach utilizing pre-trained visual language models to synthesize data similar to failure scenarios, thereby enhancing sample efficiency and planning performance for end-to-end autonomous driving systems. DriveDreamer-2 [87] converted user queries into agent trajectories via a large language model, which are then used to produce traffic-compliant HDMaps for corner case generation. Nexus [88] simulated both regular and challenging scenarios from fine-grained tokens with independent noise states to improve reactivity and goal conditioning and collected a specialized corner-case dataset to complement challenging scenario generation, Challenger [89] exploited a physics-aware multi-round trajectory refinement to identify adversarial maneuvers and a tailored scoring function to promote realistic yet challenging behaviors compatible with downstream video synthesis.

Scene Editing & Style Transfer. Many existing methods [80], [90], [91] also take world models for scene editing and style transfer to enrich the toolkit for autonomous driving simulation and data augmentation. Early methods primarily utilized scene descriptions [20] or reference images [92] for basic appearance modifications (*e.g.*, weather, lighting) and relied on bounding boxes or HD maps [73] for element-level adjustments. However, newer approaches explore richer representations for precise scene manipulation and diverse appearance control. WoVoGen [90] ensures cross-sensor consistency through world volume-aware synthesis, while SyntheOcc [75] employs occupancy grids for occlusion-aware scene editing. SimGen [91] bridges sim-to-real gaps via simulator-conditioned cascade diffusion, and

DrivePhysica [93] simulates complex driving scenarios (*e.g.*, cut-ins) using CARLA and introduces motion representation learning and instance flow guidance for temporal consistency. Complementing these, GeoDrive [94] integrates explicit 3D geometry conditions and dynamic editing to enable interactive trajectory and object manipulation.

3.1.2 Action Interpreters

Action-driven generation models bridge agent intentions and environmental dynamics through action-guided world generation and forecast-driven action planning, enabling outcome anticipation and unifying low-level maneuvers and reasoning by mapping controls to plausible futures.

Action-Guided Video Generation. Action-conditioned generation models empower agents to predict future outcomes based on intended maneuvers, effectively bridging low-level control inputs with high-fidelity video rollouts of plausible futures. GAIA-1 [60] pioneered a generative model that fuses video, text, and action inputs to synthesize realistic driving scenarios with detailed control over ego-vehicle behavior and scene attributes. GAIA-2 [126] expanded this framework to include agent configurations, environmental factors, and road semantics. GenAD [96] further enhanced generalization by releasing the OpenDV dataset alongside a predictive model that supports zero-shot, language- and action-conditioned predictions. Vista [103] applies robust action conditioning across diverse scenarios, while GEM [116] delivers multimodal outputs with precise ego-motion control, and MaskGWM [117] boosts fidelity and long-horizon predictions using mask-based diffusion. To address error accumulation in long video synthesis, InfinityDrive [107] and Epona [59] proposed memory injection and a chain-of-forward training strategy, respectively. In addition, DrivingWorld [112] generates scenarios from predefined trajectories, functioning as a neural driving simulator. Other approaches, such as DriVerse [128], MiLA [125], PosePilot [129], and LongDWM [131], focus on trajectory alignment, temporal stability, pose controllability, and depth-free guidance. Collectively, these advances drive action-conditioned generation toward better precision, temporal coherence, and robustness.

Forecasting-Driven Action Planning. Another line of work forecasts future states from current observations and ego actions, letting planners evaluate outcomes before committing [132], [133], [134]. Different from purely reactive schemes, these approaches emphasize *anticipatory decision-making*, allowing the agent to virtually "test" multiple futures and avoid unsafe trial-and-error in the real world. DriveWM [102] generates video rollouts of candidate maneuvers, scoring them with image-based rewards for trajectory selection. DriveDreamer [58] proposed the ActionFormer to predict future states and ego-environment interactions. ADriver-I [101] combines multimodal LLMs with autoregressive control signals and world evolution prediction. Vista [103] incorporates uncertainty-aware reward modules for robust action evaluation. GPT-style designs such as DrivingGPT [111] and DrivingWorld [112] model visual and action tokens jointly for planning via next-token prediction. Integrated frameworks like Doe-1 [109] unify perception, prediction, and planning for closed-loop autonomous driving, while VaVAM [122] bridges video diffusion and an action expert for decision-making. ProphetDWM [130] further couples latent action learning with state forecasting for long-term planning. Overall, by simulating diverse futures and leveraging feedback, forecast-driven models enhance generalization and safety in end-to-end autonomous driving.

3.1.3 Neural Simulators

Closed-loop simulators produce realistic virtual worlds that support effective planning, decision-making, and interaction. Regarding the difference in scene modeling, recent methods can be broadly categorized into two main approaches.

Generation-Driven Simulation. Recent advances in generative simulators for autonomous driving leverage conditional generative frameworks [8], [135], [136] to create interactive high-fidelity environments. DriveArena [119] establishes the first closed-loop framework through two core components: TrafficManager for scalable traffic synthesis and WorldDreamer for autoregressive scene generation. Building on this foundation, DreamForge [105] enhances long-term scenario modeling by integrating object-wise position encoding, supported by a novel temporal attention mechanism. Further extending these capabilities, DrivingSphere [115] introduces 4D semantic occupancy modeling that unifies static environments and dynamic objects, coupled with a visual synthesis module ensuring spatiotemporal consistency in multiview video generation. UMGen [118] simulates behavioral interactions between ego-vehicles and user-defined agents, while Nexus [88] dynamically updates environments based on agent decisions, rigorously validated through nuPlan closed-loop benchmarks. GeoDrive [94] advances trajectory optimization for VLA systems via geometry-aware scene modeling and precision control modules. Collectively, these developments transition generative simulation from passive environment rendering to closed-loop systems capable of agent interaction and feedback-driven adaptation.

Reconstruction-Centric Simulation. Reconstruction-based simulators employ neural scene reconstruction techniques such as NeRF [137] and 3D GS [138] to convert driving logs into interactive neural environments [110], [139], [140], [141], [142], [143], [144], [145], [146], [147], [148], [149], [150], [151], [152], [153]. StreetGaussian [154] represented dynamic urban street as a set of point clouds equipped with semantic logits and 3D Gaussians, each associated with either a foreground vehicle or the background. Other key implementations include HUGSIM [155], which integrates physical constraints with 3D GS for aggressive behavior synthesis, alongside frameworks like UniSim [77] and UniGaussians [156] that generate synchronized multi-modal sensor outputs through Gaussian primitive distillation. OmniRe [157] further enhances dynamic entity modeling via neural scene graph representations. While conventional 3D GS methods [21], [120], [124], [154], [158] struggle with viewpoint extrapolation artifacts, emerging solutions integrate 3D scene generation models as the data foundation to improve reconstruction robustness. ReconDreamer [159] applies progressive refinement to eliminate ghosting effects in dynamic scenes, while Stage-1 [160] achieves controllable 4D synthesis through multiview point cloud completion. These modeling methods enhanced approaches [50], [121], [159], [161], [162] demonstrate significant improvements in handling novel viewpoints, effectively bridging the fidelity gap between simulated and real-world environments.

TABLE 2: Summary of video-based generation (**VideoGen**) models.

- **Datasets:** N nuScenes [10], K KITTI [11], W Waymo Open [95], Y OpenDV-YouTube [96], A Argoverse 2 [97], N nuPlan [98], N NAVSIM [99], C CARLA [100], and P Private (Internal) Data.
- **Input & Output:** Noise Latent, Video (Single-View and/or Multi-View), and Ego-Action.
- **Architectures (Arch.):** *AR*: Autoregressive Models, *MLLM*: Multimodal Large Language Models, *SD*: Stable Diffusion Models, *DiT*: Diffusion Transformer, *GPT*: Generative Pre-trained Transformer.
- **Tasks:** *VG*: Video Generation, *E2E*: End-to-End Planning, and *3SR*: 3D Scene Reconstruction.
- **Categories:** ❶ Data Engine (Sec. 3.1.1), ❷ Action Interpreter (Sec. 3.1.2), and ❸ Neural Simulator (Sec. 3.1.3).

#	Model	Venue	Dataset	Input	Output	Condition	Len.	Freq.	Arch.	Task	Cat.	URL
1	GAIA-1 [60]	arXiv'23	P	Video	Video	T C T	L	25Hz	AR	VG	❷	
2	ADriver-I [101]	arXiv'23	N P	Video	Video Ego	T	S	2Hz	MLLM	VG	❷	
3	BEVControl [73]	arXiv'23	N	Noise	Video	3 H	S	-	SD	VG	❶	
4	BEVGen [72]	RA-L'24	N A	Video	Video	B	S	-	AR	VG	❶	
5	MagicDrive [20]	ICLR'24	N	Noise	Video	3 B C T	S	12Hz	SD	VG	❶	
6	Panacea [80]	CVPR'24	N	Video	Video	3 H C T	S	2Hz	SD	VG	❶	
7	Drive-WM [102]	CVPR'24	N	Video	Video Ego	3 H T T	S	2Hz	SD	VG, E2E	❷	
8	GenAD [96]	CVPR'24	Y	Video	Video Ego	T T	S	2Hz	SD	VG, E2E	❷	
9	DriveDreamer [58]	ECCV'24	N	Video	Video Ego	3 H T T	S	2Hz	SD	VG, E2E	❷	
10	DrivingDiffusion [81]	ECCV'24	N	Video	Video	3 T	S	2Hz	SD	VG	❶	
11	WoVoGen [90]	ECCV'24	N	Noise	Video	B T C T P	S	2Hz	SD	VG	❶	
12	Vista [103]	NeurIPS'24	Y N	Video	Video Ego	T	L	10Hz	SD	VG, E2E	❷	
13	SimGen [91]	NeurIPS'24	N	Noise	Video	3 B	S	2Hz	SD	VG	❶	
14	MagicDrive3D [21]	arXiv'24	N	Noise	Video	3 B C T	S	12Hz	SD	VG, 3SR	❶ ❸	
15	Delphi [86]	arXiv'24	N	Video	Video	3 H T	L	2Hz	SD	VG	❶	
16	BEVWorld [85]	arXiv'24	N C	Noise	Video	T	S	12Hz	AR	VG	❶	
17	Panacea+ [104]	arXiv'24	N A	Video	Video	3 H C T	S	2Hz	SD	VG	❶	
18	DiVE [76]	arXiv'24	N	Video	Video	3 H T	L	12Hz	DiT	VG	❶	
19	DreamForge [105]	arXiv'24	N	Video	Video	3 H C T	L	12Hz	SD, DiT	VG	❸	
20	SyntheOcc [75]	arXiv'24	N	Noise	Video	P	S	2Hz	SD	VG	❶	
21	HoloDrive [106]	arXiv'24	N	Video	Video	3	S	-	SD	VG	❶	
22	InfinityDrive [107]	arXiv'24	Y N	Video	Video	T T	L	10Hz	AR	VG	❷	
23	CogDriving [108]	arXiv'24	N	Video	Video	3 B	S	2Hz	DiT	VG	❶	
24	UniMLVG [92]	arXiv'24	Y N W A	Video	Video	3 B C T	L	12Hz	DiT	VG	❶	
25	DrivePhysica [93]	arXiv'24	N	Noise	Video	3 H C	L	12Hz	DiT	VG	❶	
26	Doe-1 [109]	arXiv'24	N	Noise	Video Ego	P	S	2Hz	MLLM	VG, E2E	❷ ❸	
27	OccScene [110]	arXiv'24	N K	Video	Video	T T	-	2Hz	SD	VG	❶	
28	DrivingGPT [111]	arXiv'24	N N	Video	Video Ego	T	L	10Hz	GPT	VG, E2E	❷	
29	DrivingWorld [112]	arXiv'24	P N	Video	Video Ego	T	L	10Hz	GPT	VG, E2E	❷	
30	DriveDreamer-2 [87]	AAAI'25	N	Noise	Video	3 H T	S	12Hz	SD	VG	❶	
31	SubjectDrive [82]	AAAI'25	N	Video	Video	3 H T	S	2Hz	SD	VG	❶	
32	Glad [83]	ICLR'25	N	Video	Video	3 H T	S	2Hz	SD	VG	❶	
33	DualDiff [113]	ICRA'25	N	Noise	Video	3 H C T P	S	-	SD	VG	❶	
34	DriveScape [114]	CVPR'25	N	Noise	Video	3 H T C T	S	10Hz	SD	VG	❶	
35	DriveDreamer4D [50]	CVPR'25	W	Video	Video	3 H	S	-	SD	VG, 3SR	❶ ❷	
36	DrivingSphere [115]	CVPR'25	N	Noise	Video	P	L	12Hz	DiT	VG	❸	
37	UniScene [77]	CVPR'25	N	Noise	Video	T P	L	12Hz	SD	VG	❶	
38	GEM [116]	CVPR'25	Y N	Video	Video	T	L	10Hz	SD	VG	❷	
39	MaskGWM [117]	CVPR'25	Y N W	Video	Video	T T	L	10Hz	DiT	VG	❷	
40	UMGen [118]	CVPR'25	W N	Video	Video	B T	L	2Hz	AR	VG, E2E	❸	
41	PerLDiff [79]	ICCV'25	N K	Noise	Video	3 H	S	-	SD	VG	❶	
42	DriveArena [119]	ICCV'25	N	Video	Video	3 H C T	L	12Hz	SD, DiT	VG	❶ ❸	
43	MagicDrive-V2 [74]	ICCV'25	N W	Noise	Video	3 B T C T	L	12Hz	DiT	VG	❶	
44	InfiniCube [120]	ICCV'25	W	Video	Video	H P	L	10Hz	SD	VG, 3SR	❸	
45	DiST-4D [121]	ICCV'25	N	Video	Video	3 B T C	L	12Hz	DiT	VG, 3SR	❸	
46	Epona [59]	ICCV'25	N N	Video	Video Ego	T	L	5Hz	DiT	VG, E2E	❷	
47	VaViM [122]	arXiv'25	Y	Video	Video	N/A	L	2Hz	MLLM	VG	❷	
48	VaVAM [122]	arXiv'25	Y N N	Video	Video Ego	T	L	2Hz	MLLM	VG, E2E	❷	
49	DualDiff+ [123]	arXiv'25	N	Noise	Video	3 H C T P	S	12Hz	SD	VG	❶	
50	UniFuture [124]	arXiv'25	N	Video	Video	N/A	S	12Hz	SD	VG, 3SR	❸	
51	MiLA [125]	arXiv'25	N	Video	Video	T C T	L	12Hz	DiT	VG	❷	
52	GAIA-2 [126]	arXiv'25	P	Video	Video	T C T	L	30Hz	DiT	VG	❷	
53	CoGen [127]	arXiv'25	N	Video	Video	P	L	12Hz	DiT	VG	❶	
54	Nexus [88]	arXiv'25	N W P	Video	Video	B	S	2Hz	DiT	VG	❶ ❸	
55	NoiseController [78]	arXiv'25	N	Noise	Video	3 B C T	S	12Hz	SD	VG	❶	
56	DriVerse [128]	arXiv'25	N W	Video	Video	T	L	12Hz	DiT	VG	❷	
57	PosePilot [129]	arXiv'25	N	Video	Video	C	L	2Hz	SD,DiT,AR	VG	❷	
58	GeoDrive [94]	arXiv'25	N	Video	Video	C	L	12Hz	DiT	VG, 3SR	❶ ❸	
59	Challenger [89]	arXiv'25	N	Noise	Video	3 B T	L	12Hz	DiT	VG	❶	
60	ProphetDWM [130]	arXiv'25	N	Video	Video Ego	T	L	2Hz	SD	VG, E2E	❷	
61	LongDWM [131]	arXiv'25	N	Video	Video	T	L	10Hz	DiT	VG	❷	
62	Cosmos-Drive [35]	arXiv'25	P	Video	Video	3 H T	L	-	DiT	VG	❶	
63	STAGE [84]	arXiv'25	N	Video	Video	3 H T	L	12Hz	SD	VG	❶	

3.2 World Modeling from Occupancy Generation

Generation models based on occupancy grids tailored to offer a geometry-centric representation that encodes both semantic and structural details of the 3D world. By generating, forecasting, or simulating occupancy in 3D/4D space, these models provide a geometry-consistent scaffold for perception, enable action-contingent future prediction, and support realistic large-scale simulation. Based on their primary function, existing methods can be grouped into **three** categories: ❶ **Scene Representors**, ❷ **Occupancy Forecasters**, and ❸ **Autoregressive Simulators**. Table 3 summarizes existing models under these domains.

3.2.1 Scene Representors

Occupancy-based 3D and 4D generation models, designed for learning structured 3D scene representations, treat the occupancy grid as a geometry-consistent intermediate for downstream tasks. Such a paradigm enhances perception robustness and provides structural guidance for 3D scene generation across two main applications.

3D Perception Robustness Enhancement. Occupancy-based representations have emerged as a powerful intermediate modality for enhancing perception robustness through generative modeling techniques. SSD [168] pioneered this direction by employing discrete [200] and latent diffusion [135] models for scene-level 3D categorical data generation, learning to map sparse occupancy inputs into dense semantic reconstructions. SemCity [173] further improves geometric and semantic fidelity by conditioning diffusion on initial SSC outputs, reducing inconsistencies in reconstructed scenes.

Generation Consistency Guidance. Other works leverage occupancy to guide high-fidelity, temporally coherent scene synthesis. WoVoGen [90] proposes 4D temporal occupancy volumes to drive multi-view video generation with intra-world and inter-sensor consistency. UrbanDiff [178] uses semantic occupancy grids as geometric priors for 3D-aware image synthesis, while DrivingSphere [115] transforms dynamic 4D occupancy scenes into temporally consistent video via semantic rendering. UniScene [77] generalizes occupancy-based generation across modalities, combining Gaussian-based rendering [138] with prior-guided sparse modeling for unified video and LiDAR synthesis. Collectively, these methods highlight the role of occupancy grids as a unifying structural prior for producing spatially and temporally consistent outputs with high structural fidelity.

3.2.2 Occupancy Forecasters

Models for 4D occupancy forecasting predict future occupancy from ego actions and past observations, allowing anticipation of environmental changes. This capability serves two purposes: as a self-supervised pretraining task for building generalizable 3D/4D models, and as a dynamic predictor for behavior-aware, controllable future scene generation.

Predictive Model Pretraining. Several methods explore occupancy forecasting as a pretext task to learn rich spatiotemporal features from LiDAR sequences, building generalizable generation models via self-supervised learning. EmergentOcc [63], [167] introduces differentiable rendering to reconstruct point clouds from 4D occupancy predictions, enabling self-supervised training from raw sequences. UnO [175] models a continuous 4D occupancy field for joint perception and forecasting. Large-scale pretraining frameworks such as UniWorld [169], UniScene [170], and DriveWorld [174] combine image and LiDAR data to learn foundational occupancy models that can be fine-tuned for downstream tasks like detection and planning, reducing reliance on dense labels while improving generalization.

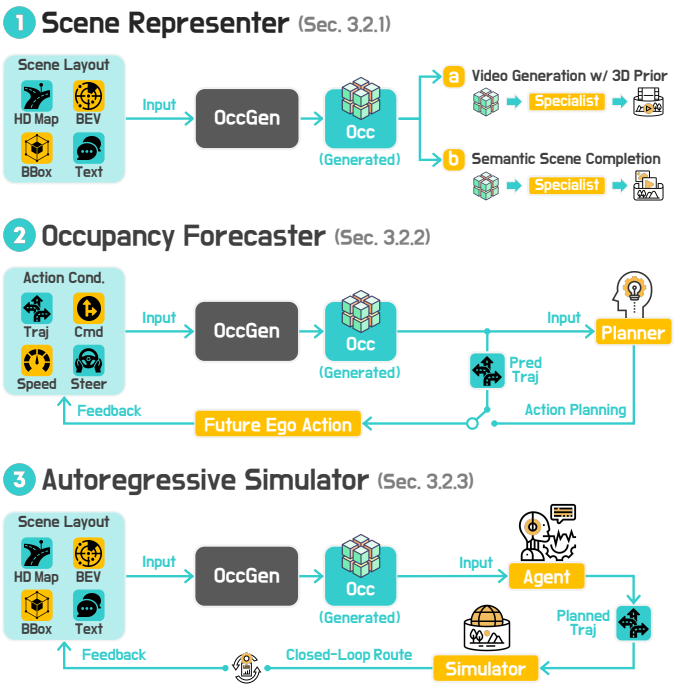

Fig. 5: The categorization of **OccGen** models based on functionalities, including scene representors (Sec. 3.2.1), forecasters (Sec. 3.2.2), and autoregressive simulators (Sec. 3.2.3).

Ego-Conditioned Occupancy Forecasting. Other approaches forecast occupancy conditioned on both history and ego-agent actions, supporting behavior-aware and controllable prediction. OccWorld [177] jointly models ego motion and surrounding environment evolution in 3D occupancy space, while OccSora [179] generates trajectory-conditioned 4D occupancy over long horizons. Later works enhance controllability [183], [197], fidelity [185], temporal coherence [182], [184], [192], and efficiency [196]. Vision-centric pipelines like Cam4DOcc [171] and its successors [186], [189] integrate world models into end-to-end planning to empower their generative abilities. OccLLaMA [181] and Occ-LLM [190] unify vision, language, and action modalities with semantic occupancy as the shared representation to support embodied question answering, while UniOcc [194] establishes a benchmark combining real and simulated data for standardized evaluation. Together, these works position occupancy forecasting as both a powerful self-supervised learning objective and a key tool for modeling dynamic, action-contingent world states.

3.2.3 Autoregressive Simulators

The occupancy-based autoregressive simulators generate large-scale, temporally coherent 4D occupancy for realistic and interactive simulation. They serve as foundation simulators for perception, planning, and decision-making,

TABLE 3: Summary of occupancy-based generation (**OccGen**) models.

- **Datasets:** S SemanticKITTI [16], C CarlaSC [163], N Occ3D-nuScenes [14], W Waymo Open [95], L Lyft-Level5 [164], A Argoverse 2 [97], 3 KITTI-360 [165], U NYUv2 [9], and O OpenCOOD [166].
- **Input & Output:** Noise Latent, Latent Codebook, Images, 3D Occ, 4D Occ, and Ego-Action.
- **Architectures (Arch.):** *Enc-Dec*: Encoder-Decoder, *LDM*: Latent Diffusion Model, *MSSM*: Memory State-Space Model, *AR*: Autoregressive Model, *DiT*: Diffusion Transformer, *LLM*: Large Language Model.
- **Tasks:** *O3G*: 3D Occupancy Generation, *O4G*: 4D Occupancy Generation, *OF*: 4D Occupancy Forecasting, *PT*: Pre-Training, *SSC*: Semantic Scene Completion, and *E2E*: End-to-End Planning.
- **Categories:** ❶ Scene Representor (Sec. 3.2.1), ❷ Occ Forecaster (Sec. 3.2.2), and ❸ Autoregressive Simulator (Sec. 3.2.3).

#	Model	Venue	Dataset	Input	Output	Condition	Len.	Arch.	Task	Cat.	URL
1	Emergent-Occ [167]	ECCV'22	N	3D Occ	3D Occ, Ego-Action	N/A	7	Enc-Dec	OF, E2E	❷	
2	FF4D [63]	CVPR'23	S N A	3D Occ	3D Occ	N/A	5	Enc-Dec	OF	❷	
3	SSD [168]	arXiv'23	C	Latent Codebook	3D Occ	N/A	1	LDM	O3G	❶	
4	UniWorld [169]	arXiv'23	N	Images	3D Occ	N/A	-	Enc-Dec	PT	❷	
5	UniScene [170]	RA-L'24	N	Images	3D Occ	N/A	-	Enc-Dec	PT	❷	
6	Cam4DOcc [171]	CVPR'24	N L	Images	4D Occ	T	4	Enc-Dec	OF	❷	
7	XCube [172]	CVPR'24	W	Noise Latent	3D Occ	N/A	1	LDM	O3G	❸	
8	SemCity [173]	CVPR'24	S C	Noise Latent	3D Occ	N/A	1	LDM	O3G, SSC	❶ ❸	
9	DriveWorld [174]	CVPR'24	N	Images	3D Occ	V S	4	MSSM	OF, PT	❷	
10	UnO [175]	CVPR'24	S N A	3D Occ	3D Occ	N/A	6	Enc-Dec	OF	❷	
11	PDD [176]	ECCV'24	S C	Latent Codebook	3D Occ	N/A	1	LDM	O3G	❸	
12	OccWorld [177]	ECCV'24	N	3D Occ	3D Occ, Ego-Action	T	6	AR	OF, E2E	❷	
13	WoVoGen [90]	ECCV'24	N	3D Occ	3D Occ	B	3	LDM	O4G	❶ ❷	
14	UrbanDiff [178]	arXiv'24	N	Latent Codebook	3D Occ	B	1	LDM	O3G	❶	
15	OccSora [179]	arXiv'24	C N W	Latent Codebook	3D Occ	T	32	DiT	O4G	❷ ❸	
16	LOPR [180]	arXiv'24	N W	3D Occ / Images	3D Occ	T	15	Enc-Dec	OF	❷	
17	OccLLaMA [181]	arXiv'24	N	Latent Codebook	3D Occ	T	6	LLM	O4G	❷	
18	FSF-Net [182]	arXiv'24	N	3D Occ	3D Occ, Ego-Action	N/A	4	Enc-Dec	OF, E2E	❷	
19	DOME [183]	arXiv'24	N	Noise Latent	3D Occ	T	11	DiT	OF	❷	
20	GaussianAD [184]	arXiv'24	N	3D Occ	3D Occ, Ego-Action	N/A	6	Enc-Dec	OF, E2E	❷	
21	OccScene [110]	arXiv'24	N S U	Images	3D Occ	N/A	1	SD	O3G	❶	
22	DFIT-OccWorld [185]	arXiv'24	N	3D Occ / Images	3D Occ, Ego-Action	T	6	Enc-Dec	OF, E2E	❷	
23	Drive-OccWorld [186]	AAAI'25	N L	Images	3D Occ, Ego-Action	C T V S	4	AR	OF, E2E	❷	
24	DynamicCity [18]	ICLR'25	C N W	Noise Latent	3D Occ	3 C T	16	DiT	O4G	❸	
25	PreWorld [187]	ICLR'25	N	Images	3D Occ, Ego-Action	T	6	Enc-Dec	OF, E2E	❷	
26	OccProphet [188]	ICLR'25	N L	Images	3D Occ	T	4	Enc-Dec	OF	❷	
27	RenderWorld [189]	ICRA'25	N	3D Occ / Images	3D Occ, Ego-Action	T	6	AR	OF, E2E	❷	
28	Occ-LLM [190]	ICRA'25	N	3D Occ / Images	3D Occ, Ego-Action	T	6	LLM	OF, E2E	❷	
29	DrivingSphere [115]	CVPR'25	N	Latent Codebook	3D Occ	B	-	LDM	O4G	❶ ❸	
30	EfficientOCF [191]	CVPR'25	N L	3D Occ	3D Occ	T	4	Enc-Dec	OF	❷	
31	UniScene [77]	CVPR'25	N	Noise Latent	3D Occ	B	6	DiT	O4G, OF	❶ ❸	
32	DIO [192]	CVPR'25	A	3D Occ	3D Occ	N/A	5	Enc-Dec	OF	❷	
33	InfiniCube [120]	ICCV'25	W	Noise Latent	3D Occ	3 H	1	LDM	O3G	❶ ❷	
34	Control-3D-Scene [193]	ICCV'25	C	Noise Latent	3D Occ	G	1	LDM	O3G	❶	
35	UniOcc [194]	ICCV'25	N C W O	3D Occ / Images	3D Occ	N/A	6	N/A	OF	❷	
36	I^2World [195]	ICCV'25	N W	3D Occ	3D Occ	C T V S	6	AR	OF	❷	
37	T^3Former [196]	arXiv'25	N	3D Occ	3D Occ, Ego-Action	T	6	AR	OF, E2E	❷	
38	COME [197]	arXiv'25	N	3D Occ	3D Occ	T	6	DiT	OF	❷	
39	$\mathcal{X}$-Scene [198]	arXiv'25	N	Noise Latent	3D Occ	3 B H	1	LDM	O3G	❶ ❸	
40	PrITTI [199]	arXiv'25	3	Noise Latent	3D Occ	3 B	1	DiT	O3G	❸	

with research focusing on two directions: generating scalable unbounded environments and modeling long-horizon dynamics for controllable closed-loop simulation.

Scalable Open-World Generation. Coarse-to-fine and outpainting strategies have been explored to construct large-scale, unbounded 3D occupancy environments. PDD [176] proposes a scale-varied diffusion framework that progressively generates outdoor scenes from coarse layouts to fine details, while XCube [172] adopts hierarchical voxel-based latent diffusion for multi-resolution generation. SemCity [173] adds manipulation functions for scene editing, and InfiniCube [120] and $\mathcal{X}$-Scene [198] integrate voxel-based occupancy with consistent visual synthesis for realistic, editable simulation worlds. Together, these works construct scalable occupancy-based representations that serve as interactive and extensible environments for embodied agents.

Long-Horizon Dynamic Simulation. Other works focus on autoregressive 4D occupancy generation to simulate dynamic world evolution. OccSora [179] produces trajectory-conditioned sequences over 16-second horizons, while DynamicCity [18] enables layout-aware and command-conditioned generation, supporting controllable scene synthesis and agent interaction. DrivingSphere [115] constructs a 4D world comprising static backgrounds and dynamic objects for closed-loop simulation, and UniScene [77] generates layout-conditioned 4D occupancy with rich semantic and geometric

detail. These approaches integrate spatial structure and temporal coherence to create realistic, controllable environments for embodied agent simulation and decision-making.

3.3 World Modeling from LiDAR Generation

LiDAR-based generation models provide geometry-aware and appearance-invariant representations by modeling complex scenes from point clouds. They enable robust 3D scene understanding and high-fidelity geometric simulation, offering advantages over image- and occupancy-based approaches in both geometric fidelity and environmental robustness. Based on their primary function, these methods can be classified into **three** categories: ❶ **Data Engines**, ❷ **Action Interpreters**, and ❸ **Autoregressive Simulators**. Table 4 summarizes existing models under these domains.

3.3.1 Data Engines

LiDAR-based data engines mitigate the scarcity of large-scale LiDAR training data due to high acquisition costs and annotation challenges by generating diverse and controllable point clouds [201], [202]. Such models enhance perception robustness, enable geometrically accurate scene completion, and support the synthesis of rare or cross-modal scenarios [49]. Recent approaches focus on four major applications.

Perception Data Augmentation. LiDAR-based generative modeling supports data augmentation for core 3D perception tasks such as detection and segmentation, with an emphasis on geometric fidelity and sensor realism. Early approaches primarily focused on modeling uncertainty and spatial structure to synthesize realistic LiDAR scans. DUSty [203] is a GAN-based framework that synthesizes realistic LiDAR scans by explicitly disentangling the underlying depth map from measurement uncertainty. DUSty v2 [204] extends DUSty by incorporating implicit neural representations, enabling the model to generate LiDAR range images at arbitrary resolutions. LiDARGen [205] pioneered the application of Langevin dynamics for LiDAR point cloud generation, achieving superior performance compared to GANs and VAEs. As the first work to adopt the denoising-diffusion paradigm in this domain, it has inspired numerous subsequent studies based on Denoising Diffusion Probabilistic Models (DDPMs) [68]. With explicit positional encoding, R2DM [206] achieves higher-precision LiDAR point cloud generation through a standardized DDPM process. Leveraging flow matching [207], R2Flow [208] significantly accelerates LiDAR point cloud generation. LiDM [209], RangeLDM [210], and 3DiSS [211] adopt latent diffusion technology by first compressing raw-scale data into low-dimensional latent variables through a pretrained VAE, then training the diffusion model in this latent space. The generated outputs are reconstructed to the original resolution, substantially improving generation speed while preserving quality. LiDARGRIT [212] extends this paradigm by discretizing the latent space with VQ-VAE [213] and generating latent codes using an autoregressive transformer. LiDARGRIT [212] further introduces a raydrop estimation loss to explicitly enhance the raydrop noise modeling. SDS [214] proposes simultaneous diffusion sampling for multi-view LiDAR scene generation, producing all views together to achieve much better geometric consistency than generating each view

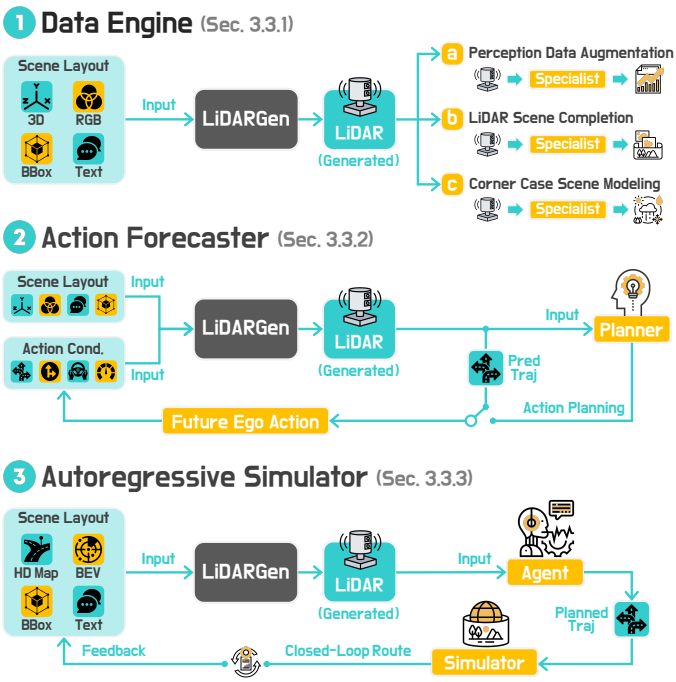

Fig. 6: The categorization of **LiDARGen** models based on functionalities, including data engines (Sec. 3.3.1), action forecasters (Sec. 3.3.2), and autoregressive simulators (Sec. 3.3.3).

separately. Recently, SPIRAL [215] pioneered the generation of segmentation-labeled LiDAR data and introduced a novel closed-loop inference strategy that enhances consistency between geometry and semantics. La La LiDAR [22] proposes a layout-guided generative framework that integrates scene graph-based layout diffusion with a foreground-aware control injector, enabling explicit modeling of object relations and controllable scene generation. Veila [216] introduces a conditional diffusion framework for panoramic LiDAR generation guided by a monocular RGB image. It addresses the challenges of reliable conditioning, cross-modal alignment, and maintaining structural coherence beyond the RGB field of view. These advances enhance LiDAR-based perception by generating diverse, controllable, and geometrically faithful training data that capture real-world sensing characteristics.

Scene Completion. The completion of 3D scenes aims to reconstruct dense and coherent 3D geometry from sparse or occluded LiDAR scans, with recent generative methods improving geometric fidelity and controllability. UltraLiDAR [220] introduces a discrete voxel-based representation for LiDAR point clouds using a VQ-VAE [213], enabling efficient and controllable sparse-to-dense completion. LiDiff [222] and DiffSSC [231] utilize the denoising process of DDPM to reposition duplicated points, thereby densifying the LiDAR point cloud while simultaneously completing occluded areas. Building on UltraLiDAR [220] for background completion and AnchorFormer [237] for foreground object synthesis, LiDAR-EDIT [228] enables flexible editing of LiDAR scenes, including object removal and insertion. By enhancing the ability to denoise large-magnitude noise, LiDPM [230] extends LiDiff [222] to generate dense point clouds not only from sparse inputs but also from pure Gaussian noise, thus enabling the synthesis of entirely novel scenes. Similarly,

TABLE 4: Summary of LiDAR-based generation (**LiDARGen**) models.

• **Datasets:** K KITTI [11], S SemanticKITTI [16], N nuScenes [10], 3 KITTI-360 [165], P PandaSet [217] C Carla [100], S SeeingThroughFog [218], W Waymo Open [95], N NAVSIM [99], A Argoverse 2 [97] and O OmniDrive-nuScenes [219].

• **Input & Output:** Noisy Latent, Latent Codebook, Noisy LiDAR Point Cloud, LiDAR Point Cloud, LiDAR Sequence, and Images/Videos (Single-View and/or Multi-View).

• **Architectures (Arch.):** *GAN*: Generative Adversarial Network, *Enc-Dec*: Encoder-Decoder, *LDM*: Latent Diffusion Model, *AR*: Autoregressive Model, *DiT*: Diffusion Transformer, *LLM*: Large Language Model.

• **Tasks:** *LG*: LiDAR Generation, *L4G*: 4D LiDAR Generation, *SEG*: 3D Semantic Segmentation, *DET*: 3D Object Detection, *SC*: Scene Completion, *OP*: Occupancy Prediction, and *E2E*: End-to-End Planning.

• **Categories:** ❶ Data Engine (Sec. 3.3.1), ❷ Action Forecaster (Sec. 3.3.2), and ❸ Autoregressive Simulator (Sec. 3.3.3).

#	Model	Venue	Dataset	Input	Output	Condition	Len.	Arch.	Task	Cat.	URL
1	DUSty [203]	IROS'21	K			N/A	1	GAN	LG, SC	❶	
2	LiDARGen [205]	ECCV'22	3 N			N/A	1	Enc-Dec	LG, SEG, SC	❶	
3	DUSty v2 [204]	WACV'23	K			L	1	GAN	LG, SEG	❶	
4	UltraLiDAR [220]	CVPR'23	S K P			P	1	Enc-Dec	LG, DET, SC	❶	
5	Copilot4D [19]	ICLR'24	A K N			T	6	LDM, AR	L4G	❶❷	
6	R2DM [206]	ICRA'24	3 K			N/A	1	Enc-Dec	LG, SC	❶	
7	ViDAR [221]	CVPR'24	N			N/A	6	Enc-Dec	L4G, DET, OP, E2E	❶❷	
8	LiDiff [222]	CVPR'24	3 S			P	1	Enc-Dec	LG, SC	❶	
9	LiDM [209]	CVPR'24	3 N S			S R T	1	LDM	LG	❶	
10	RangeLDM [210]	ECCV'24	3 N			N/A	1	LDM	LG, SC	❶	
11	Text2LiDAR [223]	ECCV'24	3 N			T	1	Enc-Dec	LG, SC	❶	
12	LiDARGRIT [212]	arXiv'24	3 K			N/A	1	Enc-Dec, AR	LG	❶	
13	BEVWorld [85]	arXiv'24	C N			T	6	LDM	L4G, DET	❶❷	
14	SDS [214]	arXiv'24	3			P	1	Enc-Dec	LG, SC	❶	
15	HoloDrive [106]	arXiv'24	N			3 T	8	LDM	L4G	❶❸	
16	LOGen [224]	arXiv'24	N			3	N/A	DiT	LG	❶	
17	OLiDM [225]	AAAI'25	3 N			T 3 O	1	Enc-Dec	LG, DET, SC	❶	
18	X-Drive [226]	ICLR'25	N			3 T	1	LDM	LG, DET	❶	
19	LidarDM [227]	ICRA'25	3 W			H	N/A	LDM	L4G, DET	❶❸	
20	LiDAR-EDIT [228]	ICRA'25	N			P	1	Enc-Dec	LG, SC, DET	❶	
21	R2Flow [208]	ICRA'25	3 N			N/A	1	DiT	LG	❶	
22	WeatherGen [229]	CVPR'25	3 S			P	1	Enc-Dec	LG, DET	❶	
23	LiDPM [230]	IV'25	S			P	1	Enc-Dec	LG, SC	❶	
24	DiffSSC [231]	IROS'25	S 3			P S	1	Enc-Dec	LG, SC	❶	
25	HERMES [232]	ICCV'25	O N			T T	N/A	AR, LLM	L4G, E2E	❶❷	
26	SuperPC [233]	CVPR'25	3			R P	1	Enc-Dec	LG, SC	❶	
27	3DiSS [211]	arXiv'25	S 3			N/A	1	LDM	LG, SEG	❶	
28	Distill-DPO [234]	arXiv'25	S			P	1	Enc-Dec	LG, SC	❶	
29	DriveX [235]	arXiv'25	N N			T	6	Enc-Dec	L4G, OP, E2E	❶❷	
30	OpenDWM [236]	arXiv'25	A 3 N W			3 H	N/A	VQ-VAE	LG, L4G	❶❸	
31	SPIRAL [215]	arXiv'25	S N			S	1	Enc-Dec	LG, SEG	❶	
32	La La LiDAR [22]	arXiv'25	N W			O	1	Enc-Dec	LG, SEG, DET, SC	❶	
33	Veila [216]	arXiv'25	K S N			R	1	Enc-Dec	LG, SEG	❶	
34	LiDARCrafter [49]	arXiv'25	N			3 O	6	Enc-Dec	L4G	❶❸	

Distillation-DPO [234] enhances both completion quality and inference efficiency of LiDiff [222] through the integration of Score Distillation [238] and Diffusion-DPO [239]. Recently, SuperPC [233] proposes a unified framework that transforms point clouds into representation features suitable for completion, upsampling, denoising, and colorization, thereby avoiding the error accumulation that can arise from sequentially applying separate models.

Rare Condition Modeling. To improve the robustness of 3D perception in adverse conditions, recent methods explore controllable LiDAR generation for safety-critical scenarios. Text2LiDAR [223] presents a Transformer-based architecture that integrates textual information to enable text-controlled LiDAR point cloud generation. WeatherGen [229] targets rainy, snowy, and foggy conditions, generating high-quality LiDAR point clouds for these conditions within a unified controllable generative model. The practical utility of the generated point cloud data is validated through 3D object detection tasks in these adverse weather scenarios. OLiDM [225] addresses fidelity limitations at the object level via a two-stage pipeline: it first generates foreground objects, which are then used as conditions for scene generation, ensuring controllable and high-quality results at both object and scene levels. Meanwhile, LOGen [224] proposes an object-level point cloud generation model to synthesize traffic participants, conditioned on their relative orientation and distance to the sensor.

Multimodal Generation. Several recent methods [85], [106] investigate multimodal generation by synthesizing aligned LiDAR and image data. X-Drive [226] introduces a dual-branch diffusion architecture for jointly generating aligned LiDAR point clouds and multi-view camera images in driving scenarios. Its key innovation is the cross-modality epipolar condition module, which improves consistency

between the point cloud and image modalities. Furthermore, X-Drive [226] supports controllable 3D scene generation conditioned on heterogeneous inputs, including text descriptions, object bounding boxes, and sensor data variants from the images or the LiDAR point clouds.

3.3.2 Action Forecasters

Based on past observations, the LiDAR-based world models functioning as action forecasters generate future LiDAR sequences conditioned on given future states.

Temporal Modeling. Copilot4D [19] proposes a scalable approach to building world models, primarily by (1) leveraging a VQ-VAE [213] model to tokenize complex, unstructured point cloud inputs, and (2) recasting the Masked Generative Image Transformer [240] as a discrete diffusion model to enable parallel denoising and decoding. Copilot4D takes as input 1–3 seconds of past LiDAR frames along with future ego actions (poses), and predicts high-quality LiDAR frames for the next 1–3 seconds. ViDAR [221] takes historical camera frames as input and predicts future LiDAR frames as output. This framework further enables pre-training for tasks such as perception, prediction, and planning.

Multi-Modal Action Forecasters. BEVWorld [85] introduces a multi-modal tokenizer to extend the generative capability to both surround-view images and LiDAR point clouds. DriveX [235] supports multi-modal outputs, including point clouds, camera images, and semantic maps. By employing a decoupled latent world modeling strategy that separates world representation learning for spatial modeling from latent future decoding for future state prediction, DriveX effectively simplifies the modeling of complex dynamics in unstructured scenes. HERMES [232] integrates LLMs to generate textual descriptions of future frames in addition to LiDAR, thereby enhancing human–machine interaction.

3.3.3 Autoregressive Simulators

World models functioning as autoregressive simulators aim to generate temporally coherent LiDAR sequences for realistic and interactive simulation. These models serve as a foundation for perception, planning, and decision-making, with a focus on geometric fidelity and temporal consistency. Existing methods can be divided into two types based on their data generation paradigms.

Sequential Autoregressive LiDAR Generation. HoloDrive [106] presents an autoregressive framework for jointly generating multi-view camera images and LiDAR point clouds by introducing a depth prediction branch in the 2D generative model to improve alignment between 2D and 3D representations. More recently, LiDARCrafter [49] extends the layout-based two-stage framework of La La LiDAR [22] to the 4D domain, with an autoregressive LiDAR sequence generator, supporting fine-grained control, long-term temporal coherence, and diverse editing capabilities.

Scene-Scale Simulation from Meshes. LidarDM [227] constructs mesh grids from point clouds by removing dynamic objects across multiple frames. It then trains a diffusion model conditioned on the BEV layout, enabling it to generate a mesh world. By incorporating dynamic objects with motion trajectories into this mesh world and performing ray projection through the scene, LidarDM can synthesize long sequential LiDAR point clouds.

4 Datasets & Evaluations

In this section, we provide a comprehensive evaluation of 3D/4D world modeling across four aspects. [1]**Datasets** (Sec.4.1) introduce widely used benchmarks with multimodal inputs and annotations across video, occupancy, and LiDAR formats. [2]**Metrics and Protocols** (Sec.4.2) define standardized criteria for assessing generation fidelity, forecasting accuracy, planning awareness, reconstruction quality, and downstream performance. [3]**Quantitative Benchmarks** (Sec.4.3) report results of state-of-the-art models under these protocols. [4]**Qualitative Analyses** (Sec.4.4) highlight strengths, limitations, and trade-offs across different modalities.

4.1 Datasets

In this survey, we discuss real, simulated, and augmented datasets that support research in 3D and 4D world modeling. These datasets span urban driving and related settings and provide rich annotations and conditions needed for **VideoGen**, **OccGen**, and **LiDARGen**. An overview of popular datasets and related benchmarks is illustrated in Figure 3. Additionally, Table 5 provides detailed statistics of each collection of the video, occupancy, LiDAR, and other relevant data formats from these mainstream datasets.

Among existing 3D/4D data collections, real-world datasets supply realism and multimodal context with reliable calibration. Recent web-scale corpora trade strict calibration for scale, diversity, and text supervision. Simulators contribute perfect labels, editable layouts, and rare or counterfactual scenarios. Together, these sources form a complementary foundation for training and evaluating controllable and planning-aware world models.

Video-based datasets provide long, coherent video sequences with reliable calibration, ego pose, and synchronized multi-view images. Conditions that aid controllability include action logs, HD maps, and language signals such as captions or driving commands. Real-world datasets, *e.g.*, nuScenes [10] and Waymo Open [95], provide surround-view imagery, accurate poses, and dense perception annotations, making them strong bases for video generation with map- or motion-conditioned control. Planning-aware datasets like NAVSIM [99] and nuPlan [98] pair short scenarios with ego motion, CAN signals, and maps to support policy-grounded video modeling. Web-scale video such as OpenDV-YouTube [96] contributes breadth and language supervision via captions and ego-action tags, trading off precise calibration for scale and diversity. Synthetic platforms like CARLA [100] offer poses and editable layouts for counterfactuals, rare events, and controlled ablations.

Occupancy-based datasets need voxelized 3D supervisions in a consistent coordinate frame, with semantic labels and tight alignment to the sensor rig. Conditions that stabilize learning include HD maps, ego trajectories, and either multi-view images or LiDAR to anchor the field over time. In driving settings, ready-to-use *real-world* benchmarks such as OpenOccupancy [242], Occ3D-nuScenes [14], NYUv2 [9], and SSCBench [243] provide standardized voxel grids and protocols for training and evaluation. *Simulated datasets* like CarlaSC [163] offer clean ground truth and full control of layout and motion, which is useful for ablations and stress tests. Semantic extensions like SemanticKITTI [16] couple

TABLE 5: Summary of datasets and benchmarks used for training **VideoGen**, **OccGen**, and **LiDARGen** models.

• **Column Keys: #** = Total number of frames; **#** = Total number of occupancy scenes; **#** = Total number of LiDAR scenes; **Freq** = Annotation frequency; Symbol "–" in a cell indicates the information is not provided.

• **Tasked by:** ❶ Video Generation Models (**VideoGen**, *cf.* Sec. 3.1), ❷ Occupancy Generation Models (**OccGen**, *cf.* Sec. 3.2), and ❸ LiDAR Generation Models (**LiDARGen**, *cf.* Sec. 3.3). Kindly refer to Table 1 for the definitions of conditions.

#	Dataset	Venue	# Scene	# (View)	#	#	Freq	Conditions	Tasked by	URL
K	KITTI [11]	CVPR'12	22	15k (×4)	-	15k	10	D 3 F	❶ ❸	
U	NYUv2 [9]	ECCV'12	464	1449 (×1)	1449	-	-	D S	❷	
C	CARLA [100]	CoRL'17	∞	∞	∞	∞	Free	3 T S	❸	
S	SemanticKITTI [16]	ICCV'19	22	-	43k	23k	10	S T	❷ ❸	
N	nuScenes [10]	CVPR'20	1000	1.4M (×6)	40k	400k	2	3 B H T V S	❶ ❷ ❸	
W	Waymo Open [95]	CVPR'20	1150	1M (×5)	-	230k	10	3 B H T V S	❶ ❷ ❸	
S	STF [218]	CVPR'20	-	1.4M (×2)	-	1.4M	0.1	3 T L P	❸	
V	Virtual KITTI 2 [241]	arXiv'20	5	40k (×2)	-	-	10	3 T	❶	
A	Argoverse 2 [97]	NeurIPS'21	1000	2.7M (×9)	-	150k	10	3 T H	❷ ❸	
L	Lyft-Level5 [164]	CoRL'21	170k	282M (×7)	42.5M	42.5M	10	T H	❷	
N	nuPlan [98]	CVPRW'21	-	24M (×6)	-	24M	-	T H	❶	
P	PandaSet [217]	ITSC'22	103	48k (×6)	-	16k	10	3 L P T S	❸	
O	OpenCOOD [166]	ICRA'22	73	11k (×4)	11k	11k	10	3 T	❷	
3	KITTI-360 [165]	TPAMI'22	379	150k (×4)	-	80k	10	3 L P T S H T	❸	
C	CarlaSC [163]	RA-L'22	24	-	43k	43k	10	T S	❷	
R	Robo3D [65]	ICCV'23	2194	-	-	476k	10	3 T S	❸	
O	OpenOccupancy [242]	ICCV'23	850	200k (×6)	34k	34k	2	3 T S	❷	
N	Occ3D-nuScenes [14]	NeurIPS'23	900	240k (×6)	40k	40k	2	3 T S	❷	
Y	OpenDV-YouTube [96]	CVPR'24	2139	60M (×1)	-	-	10	T C	❶	
S	SSCBench [243]	IROS'24	1859	404k (×6)	66k	66k	-	3 T S	❷	
N	NAVSIM [99]	NeurIPS'24	115k	920k (×8)	-	115k	2	3 T H	❸	
D	DrivingDojo [244]	NeurIPS'24	17.8k	1.7M (×1)	-	-	5	T T	❶	
O	OmniDrive [219]	CVPR'25	1000	1.4M (×6)	40k	-	2	3 T H T	❸	
E	EUVS [245]	ICCV'25	345	90k (×8)	-	-	-	T	❶	
P	Pi3DET [202]	ICCV'25	25	51k (×1)	-	51k	10	3 T	❸	

point-wise labels with occupancy volumes and enable joint learning of geometry and semantics.

LiDAR-based datasets require raw LiDAR-acquired sweeps with precise extrinsics, per-sweep ego poses, and object-level annotations. Additional 2D and 3D cues, such as HD maps, radar, and camera imagery, enable cross-modal conditioning, while coverage across weather conditions and sensor configurations improves robustness. Representative real-world sources include KITTI [11], nuScenes [10], Waymo Open [95], and Argoverse2 [97]. NAVSIM [99] supplements these with short scenario snippets paired with control signals, supporting downstream planning tasks. For robustness testing, recent benchmarks [65], [202], [218] capture adverse weather, inject systematic corruptions, and cover multiple platforms to assess generalization. Synthetic platforms, such as CARLA [100], offer clean LiDAR simulations, editable environments, and controllable signals.

4.2 Evaluation Metrics & Protocols

Standardized evaluations lay the foundation for the development of generation models. However, existing literature has overlooked the importance of establishing a systematic protocol for evaluations in 3D and 4D.

Here, we organize evaluation metrics for world models into five perspectives. [1]**Generation Quality** (Sec. 4.2.1) assesses the realism, coherence, and controllability of synthesized outputs. [2]**Forecasting Quality** (Sec. 4.2.2) evaluates future predictions given partial observations. [3]**Planning-Centric Quality** (Sec. 4.2.3) metrics measure safety and rule compliance in planning. [4]**Reconstruction-Centric Quality** (Sec. 4.2.4) examines the ability of generation models to reproduce or simulate novel views. [5]**Downstream Evaluation** (Sec. 4.2.5) tasks test how world models support tasks like detection, segmentation, and reasoning. A comprehensive summary of evaluation metrics is provided in Table 14. Together, these metrics cover both perceptual fidelity and utility in embodied decision-making and beyond.

4.2.1 *Generation Quality*

Generation quality focuses on whether a world model can produce realistic and coherent outputs given a prompt or condition. This involves four dimensions: fidelity, consistency, controllability, and human reference.

Fidelity evaluates how closely a generator matches the real data distribution and is typically divided into two families. *Perceptual metrics* project samples into a feature space learned from human-labeled data, where distances align with human judgments of realism. The Fréchet family [176], [206], [209], [224], [246], [247], [248] encodes samples, fits Gaussians to real and generated features, and reports the Fréchet distance. Some variants differ by modality and encoder, while semantic versions [215] add labels to align categories. Other representative metrics include Inception Score [249], which uses Inception logits to reward confident and diverse predictions without real references. *Statistical metrics* operate directly on geometry or density. They ask whether the generated set covers the real set, stays within it, and matches the low-level structure. Some metrics [18],

[224], [224] target the fidelity–coverage trade-off, probing set overlap by measuring whether generated samples stay on the real manifold while sufficiently covering it, while other metrics [205], [205], [212] quantify distributional discrepancy in geometry or density via different distance metrics.

Consistency evaluates whether a world model produces coherent outputs across space, time, and semantics. *Spatial Consistency* scores geometric alignment. Some [72], [102] quantify multi-view agreement by matching keypoints in overlapping regions, while others evaluate alignment by projecting the 3D outputs and comparing them with monocular depth estimates [226]. *Temporal Consistency* is measured by cosine similarity [76] between adjacent-frame embeddings from foundation models [250], [251], and *Subject Consistency* [252] tracks identity persistence by comparing subject-region features [251] across frames.

Controllability measures how well a model adheres to user-specified inputs, with metric design tailored to the conditioning modality. When the condition is reference frames, CLIP Similarity [86], [96] averages cosine similarity between CLIP embeddings of generated and reference frames to gauge semantic alignment. Beyond this, layout and object-level control is typically scored by agreement with detectors or segmentors on boxes and masks [91], scene-graph control by count errors and set overlap [193], and camera-pose control by trajectory rotation and translation errors [129],

Human Preference captures subjective qualities like realism and plausibility that automated scores may miss. Studies typically adopt either two-alternative forced choice [131] or mean opinion score [193] setups, involving both experts and lay users to provide human evaluation on world models.

4.2.2 Forecasting Quality

Forecasting quality extends beyond unconditional generation by evaluating how well the model predicts the future evolution of a scene given partial observations. Here, forecasting quality is evaluated in spatial and temporal domains.

Spatial Predictive Accuracy in forecasting measures how well predictions match the actual future in the spatial domain. For frames and videos, FID, FVD, and frame-level L1/L2 errors remain standard. IoU evaluates occupancy forecasts [171] at multiple horizons to separate near- and long-range correctness. Point-cloud forecasts [19] are evaluated by comparing the predicted and ground-truth sweeps in 3D space, using Chamfer distance for geometric overlap and depth-wise errors to quantify per-ray distance accuracy.

Temporal Predictive Accuracy in 4D forecasting assesses whether predictions remain temporally coherent, especially without full supervision [194]. Typical examples are Key Object Dimension Probability [194], which penalizes unlikely object sizes using category-specific priors, and Temporal Background Environment Consistency [194], which tracks static voxels under ego-motion to verify scene rigidity.

4.2.3 Planning-Centric Quality

Planning-centric metrics assess whether the model's outputs result in safe, efficient, and rule-compliant decisions, and its evaluation falls into open-loop and closed-loop.

Open-Loop Planning assessment evaluates predictions that do not influence future inputs. nuPlan [98] compares predictions to expert demonstrations using waypoints and heading error, and a horizon-dependent Miss Rate, which thresholds trajectory and heading errors into bounded scores. To approximate behavioral quality without full interaction, NAVSIM [99], [253] introduces short non-reactive rollouts and aggregate safety, drivable-area compliance, progress, and comfort into a single policy score, using gating and weighted averaging to align with closed-loop outcomes.

Closed-Loop Planning evaluation executes the policy in an interactive simulator and scores observed behavior. CARLA [100] reports route or goal completion and infraction distance statistics for opposite-lane driving, sidewalk incursions, and collisions with other agents. nuPlan [98] provides a broader suite of closed-loop checks, including no at-fault collisions, drivable-area and direction compliance, time-to-collision bounds, speed-limit compliance, progress along

TABLE 6: Benchmarking **VideoGen** models on the **Perceptual Fidelity** of generation quality evaluations. The reported metrics are FID and FVD scores on the official nuScenes [10] validation set. All metrics are the lower the better (↓).

Method	Resolution	Freq	FID ↓	FVD ↓
Single-View Video Generation				
DriveDreamer [58]	128×192	2 Hz	14.90	340.80
GenAD [96]	256×448	2 Hz	15.40	184.00
ProphetDWM [130]	256×448	2 Hz	6.90	190.50
Epona [59]	512×1024	5 Hz	7.50	82.80
MaskGWM [117]	288×512	10 Hz	4.00	59.40
LongDWM [131]	480×720	10 Hz	12.30	102.90
DriVerse [128]	480×832	10 Hz	18.20	95.20
InfinityDrive [107]	576×1024	10 Hz	10.93	70.06
GEM [116]	576×1024	10 Hz	10.50	158.50
Vista [103]	576×1024	10 Hz	6.90	89.40
UniFuture [124]	320×576	12 Hz	11.80	99.90
MiLA [125]	360×640	12 Hz	8.90	89.30
GeoDrive [94]	480×720	12 Hz	4.10	61.60
STAGE [84]	512×768	12 Hz	11.04	242.79
Doe-1 [109]	384×672	-	15.90	-
Multi-View Video Generation				
Drive-WM [102]	192×384	2 Hz	15.80	122.70
WoVoGen [90]	256×448	2 Hz	27.60	417.70
Panacea [80]	256×512	2 Hz	16.96	139.00
SubjectDrive [82]	256×512	2 Hz	15.98	124.00
Glad [83]	256×512	2 Hz	11.18	188.00
SynthOcc [75]	448×800	2 Hz	14.75	-
CogDriving [108]	480×720	2 Hz	15.30	37.80
DrivingDiffusion [81]	512×512	2 Hz	15.83	332.00
Delphi [86]	512×512	2 Hz	15.08	113.50
MaskGWM [117]	288×512	10 Hz	8.90	65.40
DriveScape [114]	576×1024	10 Hz	8.34	76.39
MagicDrive3D [21]	224×400	12 Hz	20.67	164.72
MagicDrive [20]	224×400	12 Hz	16.20	218.12
DreamForge [105]	224×400	12 Hz	14.61	209.90
DrivePhysica [93]	256×448	12 Hz	3.96	38.06
UniScene [77]	256×512	12 Hz	6.45	71.94
MiLA [125]	360×640	12 Hz	4.90	36.30
CoGen [127]	360×640	12 Hz	10.15	68.43
DiST-4D [121]	424×800	12 Hz	6.83	22.67
DiVE [76]	480×854	12 Hz	-	94.60
DrivingSphere [115]	480×1080	12 Hz	-	103.40
DriveDreamer-2 [87]	512×512	12 Hz	11.20	55.70
NoiseController [78]	512×1024	12 Hz	13.72	87.23
MagicDrive-V2 [74]	848×1600	12 Hz	20.91	94.84
BEVWorld [85]	-	12 Hz	19.00	154.00
UniMLVG [92]	-	12 Hz	5.80	36.10
DualDiff [113]	224×400	-	10.99	160.00
BEVGen [72]	224×400	-	24.54	-
PerLDiff [79]	256×708	-	13.36	-
HoloDrive [106]	-	-	13.60	103.00
BEVControl [73]	-	-	24.85	-

TABLE 7: Benchmarking **VideoGen** models on the **Downstream Evaluation** tasks. The reported metrics are mAP and NDS for **3D Object Detection**, mIoU (Lanes, Drivable, Divider) for **BEV Map Segmentation**, L2 and Collision Rates at timestamps 1s, 2s, and 3s for **Open-Loop Planning**, and PDMS (P) and ADS (A) scores [119] for **Closed-Loop Planning**. All results are computed using the UniAD [254] implementation and checkpoints on the official nuScenes [10] validation set.

Method	3D Det ↑		BEV Seg mIoU (%) ↑			Open-Loop Planning ↓						Closed-Loop Planning ↑			
	mAP	NDS	Lane	Dri	Div	L2@1s	L2@2s	L2@3s	CR@1s	CR@2s	CR@3s	P@SG	A@SG	P@BOS	A@BOS
Baseline [254]	37.98	49.85	31.31	69.14	25.93	0.51	0.98	1.65	0.10	0.15	0.61	-	-	-	-
MagicDrive [20]	12.92	28.36	21.95	51.46	17.10	0.57	1.14	1.95	0.10	0.25	0.70	-	-	-	-
Panacea [80]	13.72	27.73	18.23	52.37	17.21	0.58	1.14	1.95	-	-	-	-	-	-	-
DiST-4D [121]	15.63	32.44	26.80	60.32	21.69	0.56	1.11	1.91	-	-	-	-	-	-	-
DriveArena [119]	16.06	30.03	26.14	59.37	20.79	0.56	1.10	1.89	0.02	0.18	0.53	0.76	0.13	0.50	0.045
DreamForge [105]	16.63	30.57	26.16	58.98	20.22	0.55	1.08	1.85	0.08	0.27	0.81	0.81	0.12	0.74	0.076
DrivingSphere [115]	21.45	34.16	57.99	62.87	22.29	0.54	1.10	1.76	-	-	-	-	-	-	-

TABLE 8: Benchmarking **VideoGen** models on the **Downstream Evaluation** tasks. The reported metrics are mAP and NDS for **3D Object Detection** (*w/* BEVFusion [255] and StreamPETR [256]) and Road-wise mIoU scores (RmIoU) and Vehicle-wise mIoU scores (VmIoU) for **BEV Map Segmentation** (*w/* CVT [257]). The results are on the official nuScenes [10] validation set. All metrics are the higher the better (↑).

Method	BEVFusion		StreamPETR		CVT	
	mAP	NDS	mAP	NDS	RmIoU	VmIoU
Baseline	35.54	41.21	34.50	46.90	73.67	34.82
BEVControl [73]	-	-	-	-	60.80	26.80
BEVGen [72]	-	-	-	-	50.20	5.89
Panacea [80]	-	-	22.50	36.10	-	-
DrivingDiffusion [81]	-	-	-	-	63.20	31.60
SimGen [91]	-	-	-	-	62.90	31.20
CogDriving [108]	-	-	-	-	65.70	32.10
UniMLVG [92]	-	-	-	-	70.81	29.12
DrivePhysica [93]	-	-	35.50	43.67	-	-
SubjectDrive [82]	-	-	28.00	41.10	-	-
Glad [83]	-	-	27.10	40.80	-	-
DriveScape [114]	-	36.50	-	-	64.43	28.86
MagicDrive [20]	12.30	23.32	-	-	61.05	27.01
DreamForge [105]	13.01	22.16	26.00	41.10	65.27	28.36
DualDiff [113]	13.99	24.98	-	-	62.75	30.22
PerLDiff [79]	15.24	24.05	-	-	61.26	27.13
MagicDrive-V2 [74]	17.65	-	-	-	59.79	32.73
NoiseController [78]	20.93	27.96	-	-	64.85	27.32
DrivingSphere [115]	22.71	31.79	-	-	-	-

route, capturing both traffic legality and human-likeness.

4.2.4 Reconstruction-Centric Quality

Reconstruction-centric neural simulators aim to reproject the past into interactive sensor views or novel viewpoints.

Photometric Fidelity captures low-level rendering quality when ground-truth images under known viewpoints are available. Following standard practices in neural rendering, metrics such as PSNR [258], SSIM [259], and LPIPS [260] remain foundational. PSNR quantifies pixel-level accuracy, SSIM evaluates structural consistency in luminance and texture, while LPIPS measures perceptual similarity in deep feature space aligned with human visual preferences.

View Changing Consistency evaluates the spatiotemporal plausibility of novel or counterfactual viewpoints where ground truth is unavailable [50], [159]. In such settings, photometric comparison is insufficient. Metrics like Novel Trajectory Agent IoU [50] assess whether foreground agents maintain geometrically plausible behavior, offering targeted signals for validating realism in 4D interactive simulations.

4.2.5 Downstream Evaluation

While the above evaluations assess a world model in isolation, downstream evaluations measure its utility when integrated into end-to-end perception and decision-making systems. Tasks span *object detection* (mAP [261], nuScenes Detection Score [10]), *multi-object tracking* (MOTA, MOTP [262]), *semantic and BEV map segmentation* (mIoU), *3D occupancy prediction and scene completion* (voxel-level IoU, Voxelized Panoptic Quality). In language-grounded settings such as *visual question answering*, models like OccLLaMA [181] report exact-match Top-1 accuracy across question types and difficulty levels. These evaluations reflect how well a learned world model supports downstream reasoning, representation, and control tasks effectively.

4.3 Quantitative Experiments & Analyses

In this section, we quantitatively evaluate world modeling approaches through [1]**VideoGen Benchmarks** (Sec.4.3.1), [2]**OccGen Benchmarks** (Sec.4.3.2), and [3]**LiDARGen Benchmarks** (Sec. 4.3.3). Models are assessed on standardized datasets using fidelity, consistency, and forecasting metrics, along with downstream perception and planning tasks. These evaluations reveal both the progress and limitations of current methods, highlighting key trade-offs between realism, geometric accuracy, temporal stability, and controllability.

4.3.1 Benchmarking Video Generation Models

Generation Fidelity. Table 6 reports FID and FVD results on the nuScenes validation set for both single-view and multi-view vision-based world models. Early baselines such as GenAD [96] and DriveDreamer [58] operate at relatively low resolutions and frame rates, achieving modest performance (FID ∼15, FVD 180–340). Later single-view models improve visual quality. Vista [103] and InfinityDrive [107] leverage higher resolutions and frame rates, reducing FVD below 100. Recent works like MaskGWM [117] and GeoDrive [94] set new state-of-the-art, reaching FID around 4–5 and FVD near 60. In the multi-view setting, early BEV-based approaches (BEVControl [73], BEVGen [72]) yield high FID (>20). Subsequent models, including DriveWM [102], Panacea [80], and MagicDrive [20] reduce errors but struggle with temporal stability (FVD >120). Strong improvements come from models emphasizing geometric consistency and spatio-temporal alignment. UniScene [77], DriveScape [114], and DiST-4D [121] achieve the best balance, with FVD scores below 80 and DiST-4D [121] reaching as low as 22.67.

The comparison suggests resolution and frame rate strongly influence generation fidelity. Besides, explicit multi-view modeling is challenging; although many methods reduce FID, temporal coherence remains difficult, highlighting

the importance of structured 4D representations. Finally, methods combining geometry-aware priors with temporal reasoning, such as DiST-4D and UniScene, demonstrate that enforcing spatial structure and temporal consistency jointly is crucial for scalable autonomous driving video generation.

Downstream Evaluations. Table 8 and Table 7 evaluate downstream perception and planning on generated scenes. Early generative baselines (BEVControl [73], BEVGen [72]) provide limited perception benefits, especially in vehicle segmentation ($< 27\%$ mIoU). More advanced methods such as MagicDrive [20] and DreamForge [105] improve both detection (up to 26 mAP on StreamPETR [256]) and segmentation ($> 61\%$ road mIoU), while DrivePhysica [93] and Glad [83] further push detection accuracy (35.5 mAP, 43.7 NDS). For segmentation, UniMLVG [92] and CogDriving [108] achieve the highest fidelity (70.8% road, 32.1% vehicle mIoU). Beyond perception, planning performance highlights the persistent gap between synthetic and real data. While real nuScenes provides the upper bound (37.9 mAP, 49.9 NDS, 1.05 Avg L2), generative methods lag significantly in detection and planning accuracy. Nevertheless, world models like DriveArena [119] and DreamForge demonstrate reduced planning errors and collision rates, enabling preliminary closed-loop driving with non-trivial success rates (*e.g.*, 0.81 PDMS). DrivingSphere [115] achieves the strongest drivable area segmentation (>58% mIoU), while DiST-4D [121] balances detection and segmentation performance but lacks closed-loop validation.

Overall, the results show that photorealistic generation alone is insufficient to improve downstream tasks; explicit modeling of geometry, temporal consistency, and motion dynamics is crucial. Models that incorporate such priors not only enhance detection and segmentation but also support safer planning by reducing collisions and trajectory errors. Strong segmentation fidelity further demonstrates the benefit of multi-view and structure-aware models in capturing global layouts, yet the performance gap to real data remains significant, underscoring the challenge of aligning generative fidelity with task-level utility.

4.3.2 *Benchmarking Occupancy Generation Models*

Occupancy Reconstruction Quality. Table 9 evaluates the reconstruction capability of occupancy world models under VAE-based formulations. Conventional VAEs such as DOME [183] already achieve strong results (83.08% mIoU, 77.25% IoU), outperforming most VQVAEs. While UrbanDiff [178] and I^{2}World [195] show competitive IoU, other variants like OccSora [179] degrade significantly under coarse temporal–spatial compression. Triplane-based VAEs [173], [196], [198] bring the largest gains, with T^{3}Former [196] reaching 85.50% mIoU and X-Scene [198] establishing a new state-of-the-art at 92.40% mIoU and 85.60% IoU.

These results underline that latent representation design is decisive for reconstruction fidelity. Triplane factorization enforces geometric consistency and enables finer spatial detail, while simply enlarging latent dimensionality (*e.g.*, UrbanDiff [178] with 2048 channels) yields limited returns. Compact VAEs such as UniScene [77] further show that well-regularized low-dimensional spaces can generalize effectively, whereas aggressive compression (*e.g.*, OccSora [179]) sacrifices accuracy. Overall, effective compression combined with explicit geometric priors is key to scalable and accurate 3D and 4D scene modeling.

TABLE 9: Benchmarking **OccGen** models on **Reconstruction Quality**. The reported metrics are mIoU (%) for **Semantic Occupancy Reconstruction** and IoU (%) for **Occupancy Reconstruction**. All results are on the official nuScenes [10] validation set. Both metrics are the higher the better (↑).

Method	Type	Resolution	mIoU ↑	IoU ↑
OccSora [179]	VQVAE	$(\frac{T}{8}, 25, 25, 512)$	27.40	37.00
OccLLaMA [181]	VQVAE	$(50, 50, 128)$	65.93	57.66
OccWorld [177]	VQVAE	$(50, 50, 128)$	66.38	62.29
UrbanDiff [178]	VQVAE	$(50, 50, 2048)$	80.00	98.80
I^{2}World [195]	VQVAE	$(50, 50, 128)$	81.22	68.30
Occ-LLM [190]	VAE	$(50, 50, 64)$	71.08	62.74
UniScene [77]	VAE	$(50, 50, 8)$	72.90	64.10
DOME [183]	VAE	$(25, 25, 64)$	83.08	77.25
UniScene [77]	VAE	$(100, 100, 8)$	92.10	87.00
T^{3}Former [196]	Triplane-VAE	$(100, 100, 16, 8)$	85.50	72.07
X-Scene [198]	Triplane-VAE	$(100, 100, 16, 8)$	92.40	85.60

TABLE 10: Benchmarking **OccGen** models on **4D Occupancy Forecasting Quality**. The reported metrics are mIoU (%) for **Semantic Occupancy Reconstruction** and IoU (%) for **Occupancy Reconstruction**, respectively, at timestamps 1s, 2s, and 3s. All results are on the official nuScenes [10] validation set. Both metrics are the higher the better (↑).

Method	mIoU (%) ↑			IoU (%) ↑		
	1s	2s	3s	1s	2s	3s
GaussianAD [184]	6.29	5.36	4.58	14.13	14.09	14.04
PreWorld [187]	12.27	9.24	7.15	23.62	21.62	19.63
Occ-LLM [190]	24.02	21.65	17.29	36.65	32.14	28.77
OccLLaMA [181]	25.05	19.49	15.26	34.56	28.53	24.41
OccWorld [177]	25.78	15.14	10.51	34.63	25.07	20.18
RenderWorld [189]	28.69	18.89	14.83	37.74	28.41	24.08
COME [197]	30.57	19.91	13.38	36.96	28.26	21.86
DFIT-OccWorld [185]	31.68	21.29	15.18	40.28	31.24	25.29
DOME [183]	35.11	25.89	20.29	43.99	35.36	29.74
UniScene [77]	35.37	29.59	25.08	38.34	32.70	29.09
T^{3}Former [196]	46.32	33.23	28.73	77.00	75.89	76.32
I^{2}World [195]	47.62	38.58	32.98	54.29	49.43	45.69

4D Occupancy Forecasting Quality. Table 10 presents 4D occupancy forecasting results over the period of 1–3 seconds. Baselines such as OccWorld [177] and OccLLaMA [181] achieve moderate performance (17–20% mIoU), while DOME [183] and UniScene [77] improve temporal stability (27.10% and 31.76% mIoU). More recent models show further progress: I^{2}World [195] reaches 39.73% mIoU with balanced IoU, and T^{3}Former [196] excels in spatial coherence with 76.40% IoU. The comparisons reveal three insights. First, naive autoregressive or generative approaches deteriorate rapidly at longer horizons, highlighting the need for structured priors. Second, triplane factorization substantially improves spatial fidelity, as reflected in the performance of T^{3}Former [196]. Third, I^{2}World shows that coupling scalable latent reasoning with temporal modeling yields the best balance across horizons. Accurate 4D forecasting thus requires not only generative power but also structured representations that enforce geometric and temporal consistency.

End-to-End Planning. Table 11 reports the performance of end-to-end planning, measured by trajectory error (L2) and collision rate. Sequence-based planners like ST-P3 [263] perform poorly (2.11 meters in L2 error), while UniAD [254] and GenAD [96] achieve substantial gains,

Fig. 7: Qualitative comparisons of state-of-the-art **VideoGen** models on the nuScenes [10] dataset. From top to bottom rows: Reference (from the dataset), MagicDrive [20], DreamForge [105], DriveDreamer-2 [87], and OpenDWM [236].

TABLE 11: Benchmarking **OccGen** models on **Motion Planning Quality**. The reported metrics are L2 Error Rate (in meters) and Collision Rate (%), respectively, at timestamps 1s, 2s, and 3s. All results are on the official nuScenes [10] validation set. Both metrics are the lower the better (↓).

Method	L2 Error (m) ↓			Collision Rate (%) ↓		
	1s	2s	3s	1s	2s	3s
ST-P3 [263]	1.33	2.11	2.90	0.23	0.62	1.27
OccNet [264]	1.29	2.13	2.99	0.21	0.59	1.37
FSF-Net [182]	0.54	1.09	-	0.01	0.01	-
UniAD [254]	0.48	0.96	1.65	0.05	0.17	0.71
OccWorld [177]	0.43	1.08	1.99	0.07	0.38	1.35
PreWorld [187]	0.41	1.16	2.32	0.50	0.88	2.42
GaussianAD [184]	0.40	0.64	0.88	0.09	0.38	0.81
DFIT-OccWorld [185]	0.38	0.96	1.75	0.07	0.39	0.90
Occ-LLaMA [181]	0.37	1.02	2.03	0.04	0.24	1.20
GenAD [96]	0.36	0.83	1.55	0.06	0.23	1.00
RenderWorld [189]	0.35	0.91	1.84	0.05	0.40	1.39
T^3Former [196]	0.32	0.91	1.76	0.08	0.32	0.51
Drive-OccWorld [186]	0.32	0.75	1.49	0.05	0.17	0.64
Occ-LLM [190]	0.12	0.24	0.49	-	-	-

with UniAD+DriveWorld [174] further improving to 0.69 meter in L2 error and 0.19% collisions. Occupancy-based world models such as OccWorld [177] and OccLLaMA [181] reduce errors to around 1.15 meters. Structured refinements (*e.g.*, DFIT-OccWorld [185], RenderWorld [189], and Drive-OccWorld [186]) achieve stronger accuracy and safety, with Drive-OccWorld reaching 0.85 m in L2 error and 0.29% collisions. Notably, GaussianAD [184] and T^3Former [196] balance error and safety, while Occ-LLM [190] reports extremely low error (*i.e.*, only 0.28 meter in L2 error). The results show that integrating occupancy world models into planning pipelines consistently outperforms pure trajectory-based methods. Hybrid designs that refine occupancy priors, such as Drive-OccWorld [186] and DFIT-OccWorld [185], bring joint improvements in accuracy and safety, demonstrating the downstream robustness of generative modeling. Overall, structured occupancy representations form a strong foundation for end-to-end autonomous driving, enabling reliable long-horizon planning in complex scenarios.

4.3.3 *Benchmarking LiDAR Generation Models*

Generation Fidelity. Table 12 reports the performance of recent LiDAR scene generation methods on SemanticKITTI [16] using four fidelity metrics (FRD, FPD, JSD, and MMD). Earlier methods such as LiDARGen [205] and LiDM [209] exhibit relatively large distributional discrepancies, as reflected by high FRD and FPD scores. In contrast, more recent approaches, including R2DM [206], Text2LiDAR [223], and WeatherGen [229], achieve substantially better results across most metrics, indicating a closer alignment between generated and real LiDAR distributions.

The results reveal a clear progression in LiDAR generation quality. Among evaluated methods, WeatherGen [229] achieves the best performance across all metrics by employing Mamba [265] as its backbone. Interestingly, Text2LiDAR [223], despite its strong conditioning on textual input, produces higher FRD, suggesting that aligning with semantic prompts may compromise geometric fidelity. These

Fig. 8: Qualitative comparisons of state-of-the-art **VideoGen** models on the nuScenes [10] dataset. From top to bottom rows: Reference (from the dataset), MagicDrive [20], DreamForge [105], DriveDreamer-2 [87], and OpenDWM [236].

TABLE 12: Benchmarking **LiDARGen** models on the **Perceptual Fidelity** evaluations. The reported metrics are FRD, FPD, JSD and MMD scores on the official SemanticKITTI [16] dataset. All metrics are the lower the better (↓).

Method	Resolution	FRD↓	FPD↓	JSD↓	MMD↓
LiDARGen [205]	64×1024	681.37	115.17	0.1323	2.19×10^{-3}
LiDM [209]	64×1024	-	108.70	0.0456	2.90×10^{-4}
R2DM [206]	64×1024	192.81	19.29	0.0373	1.60×10^{-4}
Text2LiDAR [223]	64×1024	522.32	11.09	0.0750	4.29×10^{-4}
WeatherGen [229]	64×1024	184.11	11.42	0.0290	3.80×10^{-5}

TABLE 13: Benchmarking **LiDARGen** models on **4D LiDAR Generation Quality**. The reported metrics are TTCE and CTC. The numbers indicate frame intervals. All results are on nuScenes [10]. Both metrics are the lower the better (↓).

Method	TTCE ↓		CTC ↓			
	3	4	1	2	3	4
UniScene [77]	2.74	3.69	0.90	1.84	3.64	3.90
OpenDWM [236]	2.68	3.65	1.02	2.02	3.37	5.05
OpenDWM-DiT [236]	2.71	3.66	0.89	1.79	3.06	4.64
LiDARCrafter [49]	2.65	3.56	1.12	2.38	3.02	4.81

findings underscore the importance of balancing semantic controllability with distributional realism in future LiDAR scene generation research.

4D LiDAR Generation Quality. Table 13 benchmarks recent LiDAR-based 4D scene generation methods on temporal coherence, using TTCE (Temporal Transformation Consistency Error) and CTC (Chamfer Temporal Consistency) as evaluation metrics. Unlike video generation, which has been extensively studied with standardized benchmarks, temporal LiDAR generation remains relatively underexplored, and current metrics mainly focus on explicit geometric alignment across frames. The results reveal several observations. First, end-to-end autoregressive methods such as UniScene [77] and OpenDWM-DiT [236] demonstrate clear advantages in maintaining short-horizon geometric consistency, as reflected in lower TTCE and CTC at 1–2 frame intervals. However, their fixed-length generation limits broader applicability, as error accumulation grows at longer horizons. Second, incorporating strong vector quantization modules [236] facilitates better condition embedding and fine-grained reconstruction, leading to improved temporal stability. Third, modality choices introduce inherent trade-offs: BEV-based generation offers smoother temporal continuity but sacrifices fidelity to the raw point cloud pattern, while range-based [49] generation better preserves LiDAR-specific sensing characteristics but requires careful design to embed conditions and sustain long-term consistency.

4.4 Qualitative Experiments & Analyses

In this section, we qualitatively evaluate the 3D and 4D generation approaches through [1]**VideoGen Visualizations** (Sec.4.4.1), [2]**OccGen Visualizations** (Sec.4.4.2), and [3]**LiDARGen Visualizations** (Sec. 4.4.3). These evaluations highlight the strengths, limitations, and trade-offs of current methods, informing future advances in realism, consistency, and generalization.

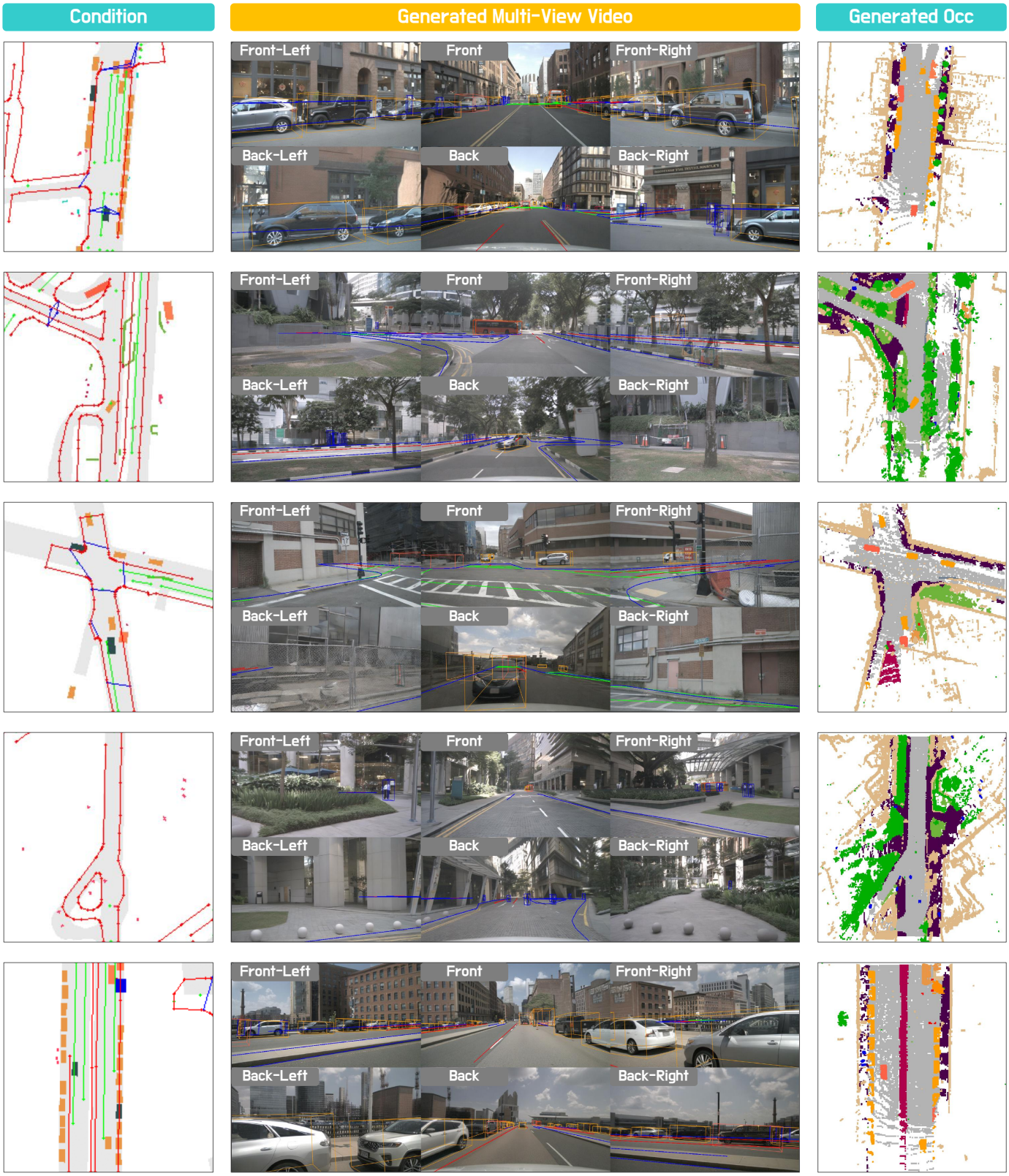

Fig. 9: Qualitative examples of **OccGen** models on the nuScenes [10] dataset. From left to right columns: The input condition, the generated multi-view videos, and the generated occupancy grids. The results are generated using $\mathcal{X}$-Scene [198].

4.4.1 Qualitative Analyses of VideoGen Models

Visual Realism. Figure 7 and Figure 8 compare recent video generation world models, including MagicDrive [20], DreamForge [105], DriveDreamer-2 [58], and OpenDWM [236]. The generated scenes capture overall layouts and semantics close to real-world distributions, but fine-grained details often suffer from pixel misalignment, blurred textures, and structural discontinuities. Among the methods, OpenDWM [236] achieves the most realistic, consistent, and controllable results, owing to its training on diverse datasets (OpenDV [96], nuScenes [10], and Waymo Open [95]), while others rely on a single dataset. This underscores the role of dataset diversity in improving generalization and robustness.

Physical Plausibility. In the absence of explicit physics constraints, generated videos may exhibit violations of physical realism, such as vehicle–background interpenetration, incorrect shadows, or scale distortions. While such issues may appear subtle in static frames, they significantly reduce realism when viewed as continuous video, undermining temporal coherence and physical plausibility.

Controllability. Appearance-level controls (weather, time-of-day, style) can be reliably controlled via large-scale pretrained video generation models with text conditioning. By contrast, precise geometric control over object position, orientation, and velocity remains challenging, typically requiring dedicated control embeddings or structured conditioning mechanisms.

Long-Tail Categories. Rare and small-scale classes (e.g., pedestrians, cyclists, traffic signs) remain hard to generate convincingly. Long-tail data imbalance often leads to unrealistic shapes, distorted geometry, or even omission of these critical objects.

Takeaways. The results suggest that future progress in video-based world models requires advances along five critical axes: (i) **realism**, reducing artifacts and enhancing detail fidelity; (ii) **consistency**, maintaining semantic and temporal coherence; (iii) **controllability**, unifying high-level appearance control with fine-grained geometric control; (iv) **physical plausibility**, incorporating physics priors to prevent unrealistic artifacts; and (v) **generalization**, leveraging diverse large-scale datasets to improve robustness.

4.4.2 Qualitative Analyses of OccGen Models

3D Geometric Consistency. Figure 9 shows qualitative results of occupancy generation models conditioned on scene layouts. The generated multi-view videos and occupancies exhibit strong spatial alignment across different perspectives. Such cross-view coherence is crucial for maintaining geometric plausibility in multi-camera settings.

Occupancy Fidelity. The generated occupancies preserve key semantics, including drivable areas, sidewalks, and surrounding objects. While overall layouts are captured reliably for downstream perception, fine-grained geometry (e.g., thin lane boundaries, small dynamic agents) remains challenging, often leading to misalignment or incomplete reconstruction.

Controllability and Generalization. Conditioned on high-level scene priors, models can flexibly adapt to diverse intersection layouts and road structures, demonstrating promising controllability. However, rare structures and long-tail categories (e.g., bicycles, pedestrians) are often poorly represented, revealing limitations in data diversity and generalization capacity.

Takeaways. These results suggest that progress in occupancy generation hinges on three aspects: (i) **geometric consistency**, ensuring spatial coherence across 3D environments; (ii) **fine-grained fidelity**, particularly for small-scale and dynamic objects; and (iii) **generalization**, leveraging diverse datasets to handle rare layouts and long-tail classes. Advancing these aspects is essential for robust world models capable of supporting downstream tasks and closed-loop simulation.

4.4.3 Qualitative Analyses of LiDARGen Models

Global Patterns. Figure 10 compares representative LiDAR generation paradigms. The original scans exhibit dense rings with uniform angular spacing, faithfully capturing both static structures and dynamic objects. The voxel-based OpenDWM [236] emphasizes coherent scene geometry but often yields overly regularized patterns due to voxel-level modeling. The range-based LiDARCrafter [49] better preserves the native scanline structure with sharper rings, though it may introduce artifacts around occlusion boundaries. The occupancy-based UniScene [77] reproduces global distributions but tends to oversmooth fine details, leading to discontinuities.

Point Cloud Sparsity. Given the inherent sparsity of LiDAR data, generation models must balance realistic density with structural consistency. OpenDWM [236] often produces overly sparse regions, especially at long ranges. LiDARCrafter [49] maintains more uniform angular density, closely following the sensor's scanning characteristics. UniScene [77] provides globally complete coverage but sometimes introduces artificial filling inconsistent with real sensor patterns.

Object Completeness. Dynamic agents such as vehicles are particularly important for downstream perception and planning. OpenDWM [236] frequently underrepresents object contours, resulting in fragmented or partial shapes. LiDARCrafter [49] offers better surface completion, though finer details can be noisy. UniScene [77] reconstructs volumetrically plausible objects with consistent occupancy, but often lacks the sharp boundaries and crisp detail of real scans.

Takeaways. These results highlight three key attributes for LiDAR generation: (i) **global patterns**, ensuring coherent scene geometry while preserving sensor-specific scan structures; (ii) **point sparsity**, maintaining realistic density distributions that match LiDAR characteristics; and (iii) **object completeness**, accurately capturing dynamic agents with sharp contours and consistent surfaces. Future advances will require balancing these attributes to generate LiDAR sequences that are both perceptually realistic and physically faithful to sensor properties.

5 Applications

The versatility of 3D and 4D world models enables deployment across diverse domains. [1]**Autonomous Driving** (Sec. 5.1) supports simulation, evaluation, and scenario synthesis. [2]**Robotics** (Sec. 5.2) leverages them for navigation, manipulation, and scalable simulation. [3]**Video Games & XR** (Sec. 5.3) benefit from content generation, immersive rendering, and adaptive environments. [4]**Digital Twins** (Sec. 5.4) enable city-scale reconstruction, event replay, and scene editing.

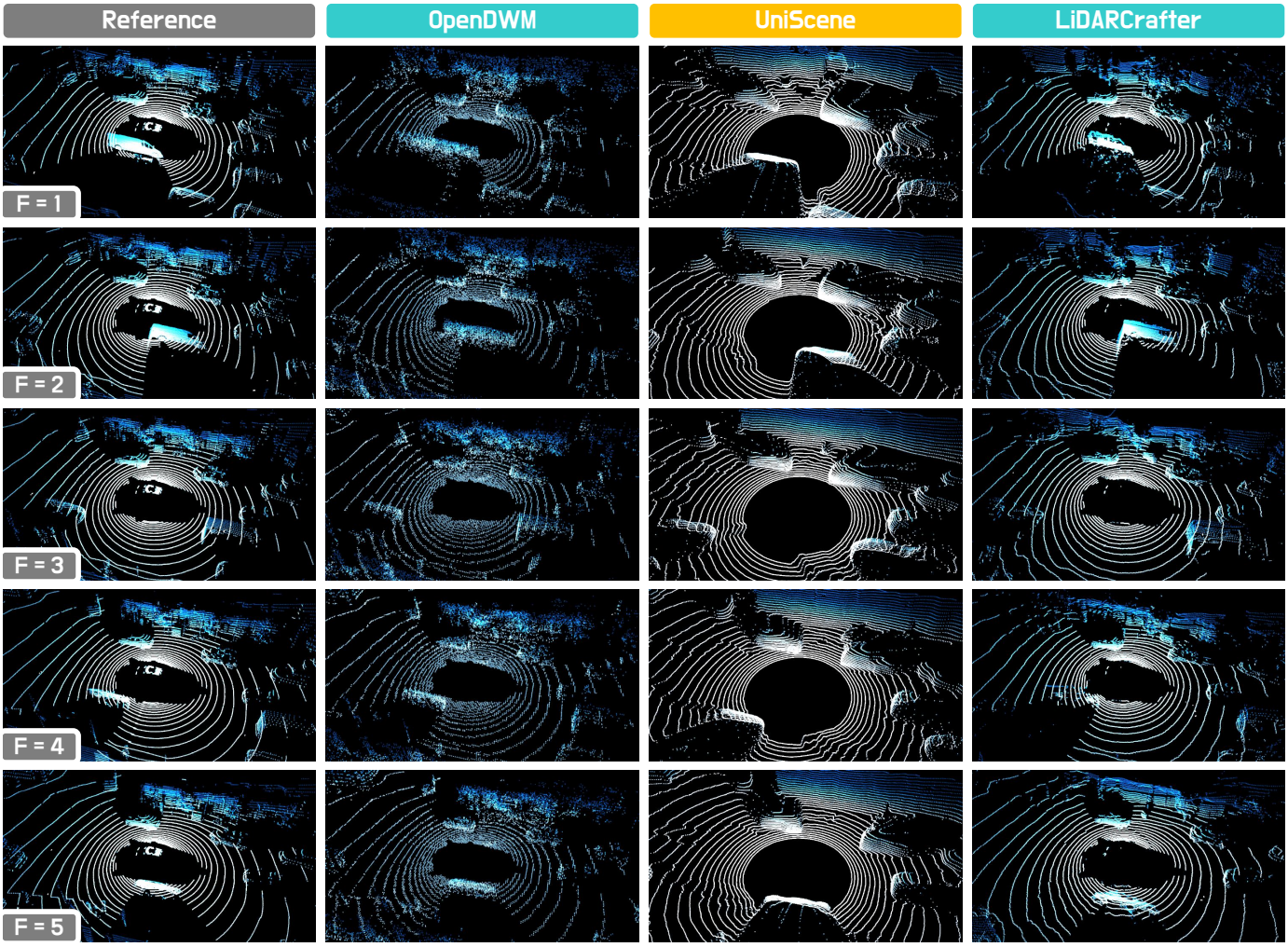

Fig. 10: Qualitative comparisons of state-of-the-art **LiDARGen** models on the nuScenes [10] dataset. From left to right columns: Reference (from the dataset), OpenDWM [236], UniScene [77], and LiDARCrafter [49].

[5]**Emerging Applications** (Sec. 5.5) span scientific discovery, healthcare, industry, and disaster response. Together, these applications showcase the role of world models in unifying perception, prediction, and generation across domains.

5.1 Autonomous Driving

3D and 4D world models provide a principled foundation for autonomous driving, supporting simulation, evaluation, and scenario synthesis. They enable controllable, interactive, and safety-critical environments that cannot be easily reproduced in the real world. We outline three major applications.

Traffic Simulation. World models enable realistic traffic simulators with heterogeneous agents, diverse motion, and physics-compliant interactions. Compared with image-only platforms, volumetric representations such as occupancy grids [179], [185], multi-frame LiDAR point clouds [49], or scene-level meshes [12] provide richer geometry and temporal coherence [102], [266]. Modern systems further support controllable parameters (*e.g.*, traffic density, intent, weather) and stochastic perturbations, improving robustness and generalization for downstream policies [58], [60], [80], [87], [104], [125], [126], [267].

Closed-Loop Driving Evaluation. Beyond static benchmarks, closed-loop setups couple generative models with agents to assess *perception→planning→control* stacks over long horizons [50], [266]. By jointly modeling ego behavior and surrounding traffic dynamics, models create responsive environments that adapt to agent actions in real time [49], [94]. This allows scalable evaluation of robustness under distribution shifts, rare events, and recovery after failures [126], [267], while modular conditioning (*e.g.*, HD maps, text queries, and ego trajectories) enables targeted stress testing [74], [94].

Scenario Synthesis. World models can generate rare or safety-critical driving scenes that are underrepresented in real datasets, which is essential for evaluating robustness. Typical cases include severe occlusions, sudden intrusions, multi-agent conflicts, and adverse weather [22], [215], [268]. Controllable generation with HD maps, semantic masks, scene graphs, or textual prompts enables targeted testing [49], [94], [215]. Physics- and motion-aware models ensure dynamic feasibility [269], [270], while stochastic sampling improves coverage of rare events. LiDAR-centric approaches such as LiDARCrafter [49] further extend this capability to 4D sequences with temporal coherence.

5.2 Robotics

3D and 4D world models have the potential to enhance robotic intelligence by supporting navigation, manipulation, and simulation. They provide spatial-temporal grounding, physical reasoning, and scalable synthetic environments, which are crucial for robust policy learning.

Embodied Navigation. Robots leverage world models to perceive and predict dynamic layouts, enabling long-horizon exploration, obstacle avoidance, and localization in both structured and unstructured settings [25], [271], [272]. Forecasting future states is critical in crowded or occluded scenes [271], [273], where multi-scan LiDAR, voxelized occupancy, and predictive dynamics provide reliable spatial-temporal cues [266], [274]. Recent studies also combine visual, topological, and linguistic signals for instruction following and adaptive decision-making [25], [275].

Object-Centric Manipulation. For this task, models capture object geometry and physical transitions, allowing robots to anticipate contact dynamics and plan stable grasps or rearrangements [26], [276], [277]. Representations such as meshes, keypoint graphs, and volumetric embeddings support fine-grained control and generalization to new objects [278], [279]. Integration of differentiable physics with generative models yields physically consistent predictions that can be optimized for various tasks [26], [277], [280].

Scene Generation for Simulation. Generative models create diverse synthetic environments, reducing manual design costs for training and evaluation [271], [281], [282]. Procedural variation in layout, semantics, and dynamics exposes robots to a wide range of scenarios, improving robustness and sim-to-real transfer [25], [26], [272], [280]. Flexible scene representations, from meshes to voxel grids and point clouds, further enable integration with both physics-based simulators and photorealistic renderers [277], [278].

5.3 Video Games & XR

World models transform gaming and XR by automating content creation, supporting immersive rendering, and enabling adaptive environments that respond to player actions.

Procedural World Generation. Generative models automate the design of expansive virtual worlds, supporting open exploration and emergent gameplay [283], [284], [285]. Procedural pipelines can incorporate maps, player states, or language prompts to scale content production beyond manual asset creation [283], [286]. Maintaining temporal and semantic coherence is key for believable dynamic evolution [287], while diverse scene representations such as point clouds, voxels, and neural radiance fields balance realism, style, and efficiency [138], [213].

Interactive Scene Rendering. Immersive XR requires real-time rendering of dynamic scenes where users move freely through evolving geometry and lighting [288], [289]. Neural representations including NeRF [213] and Gaussian Splatting [138] advance photorealistic synthesis, with temporal extensions modeling motion and state change [290], [291]. To ensure consistency and comfort, systems must maintain geometric fidelity under arbitrary viewpoints, adapt scene content to user actions [29], [292], and employ efficient pipelines to sustain high frame rates.

Playable Environment Adaptation. Adaptive worlds adjust geometry, layout, and agent behavior to sustain challenge and engagement [284], [293], [294], [295]. 3D/4D models support real-time transformations such as altering terrain, collapsing structures, or spawning entities based on player interactions [286], [293]. By leveraging priors or high-level instructions, these systems preserve style, physics, and narrative coherence [49], [287], thereby enhancing immersion, replayability, and personalized gameplay.

5.4 Digital Twins

3D and 4D world models underpin urban digital twins by enabling large-scale reconstruction, event replay, and interactive editing. These capabilities support planning, analysis, and simulation in smart city applications.

City-Scale Scene Modeling. Digital twins integrate multimodal sensing, including LiDAR, RGB-D, aerial photogrammetry, and drone surveys, to capture both static infrastructure and dynamic activities [18], [296], [297]. They enable applications such as traffic monitoring, infrastructure planning, and disaster response [16], [298], while dynamic modeling simulates pedestrian and vehicle flows for capacity planning [299], [300]. Recent advances in streaming pipelines and 4D compression maintain temporal consistency and allow metropolitan-scale deployment [5], [301].

Event Replay & Forecasting. World models reconstruct past or hypothetical events from sparse sensor logs, aiding analysis of incidents [302], [303], construction monitoring [304], or emergency response [305]. Replayable 4D scenes clarify causality, while predictive extensions enable what-if simulations for evaluating interventions. Alignment with sensor ground truth remains critical for reliability.

Scene Control & Editing. Interactive tools allow users to manipulate urban content for simulation and visualization, including vehicle removal, weather alteration, and layout modification [287], [301]. Such controllability improves planning workflows and supports immersive city-scale analysis.

5.5 Other Emerging Applications

Beyond autonomous driving and robotics, 3D and 4D world models are expanding into scientific, medical, industrial, and safety-critical domains. These applications highlight their versatility in modeling complex spatial–temporal systems.

Scientific Discovery and Environmental Modeling. World models capture natural dynamics from multimodal observations, supporting forecasting and exploratory simulation. Applications include climate and weather prediction [306], [307], [308], monitoring glacier retreat or floods, and simulating wildfire spread. By learning directly from data, they complement physics-based solvers with faster iteration.

Healthcare & Biomechanics. Generative 3D models reproduce anatomy deformation and tissue behavior for surgical training, planning, and guidance [309]. Predictive motion models aid rehabilitation, prosthetics, and injury prevention by anticipating joint trajectories [310], enhanced by multi-view capture and volumetric reconstruction.

Industrial Process & Manufacturing Simulation. Virtual prototyping with world models supports robotic assembly, material handling, and inspection [157], [311]. Temporal

simulation of component interactions reduces costly trials and enables analysis of efficiency and fault recovery.

Security, Defense & Disaster Response. Synthetic environments simulate tactical operations, hazardous conditions, and evacuations [312]. Dynamic scene modeling further aids disaster preparedness by predicting structural collapse, fire spread, or chemical dispersion, and testing emergency response plans.

6 Challenges & Future Directions

In this section, we highlight key challenges of world models, including benchmarking, long-horizon fidelity, physical realism, efficiency, and cross-modal coherence, and outline directions for future research.

6.1 Standardized Benchmarking & Evaluations

A major barrier to progress in the driving world models is the lack of common, standardized benchmarks and evaluation protocols. Current studies often utilize different datasets or ad hoc metrics, which makes it difficult to meaningfully compare models and assess their true performance in diverse realistic settings [252], [313], [314], [315]. Establishing unified benchmarks can provide a comprehensive evaluation framework that captures key metrics such as physical plausibility, temporal consistency, and controllability. Moreover, standardized evaluations should encompass both closed-loop simulation tests and real-world scenarios to validate the model's capabilities under varying traffic densities, weather conditions, and complex urban architectures [119]. Future work must focus on developing these benchmarks to ensure fair and transparent comparisons across different approaches.

6.2 High-Fidelity & Long-Horizon Generation

Another critical challenge in world models for autonomous driving is achieving high-fidelity generation over long time horizons [92], [105]. While short-term predictions may capture immediate interactions with reasonable accuracy, small errors tend to accumulate over longer sequences, leading to unrealistic behaviors and degradation of scene consistency. The difficulty of maintaining both high visual fidelity and long-horizon coherence is compounded by the complexity of dynamic urban environments, where interactions between multiple agents and environmental factors evolve continuously. Addressing these issues requires advanced generative techniques that explore novel training paradigms [59] and memory mechanisms [107] that effectively penalize long-term divergences to enable reliable long-term simulation.

6.3 Physical Fidelity, Controllability & Generalizability

From the perspective of the generation capability, current world models for autonomous driving are critically limited by a failure to ensure *physical realism*, offer fine-grained controllability, and achieve robust generalization [74], [93]. They often produce physically implausible events, such as non-deforming collision impacts and objects that lack temporal consistency [47]. Furthermore, their editing capabilities remain coarse, typically confined to adjusting traffic agents' positions or appearances while neglecting granular control over environmental elements like architecture or road signs. Most critically, these models tend to overfit their training data, failing to generalize to new urban environments and rare objects, thus limiting their real-world applicability. Future work must overcome these challenges to build more faithful, controllable, and generalizable world models.

6.4 Computational Efficiency & Real-Time Performance

Another pressing limitation of current world models for autonomous driving lies in computational efficiency and real-time responsiveness. Existing methods often depend on heavy architectures and multi-step sampling strategies, leading to substantial latency and memory overhead, which undermines their practicality for large-scale data generation and simulation. Moving forward, research should prioritize sparse computation [316] and inference acceleration techniques [317] in order to enable world models that are both accurate and responsive while remaining scalable.

6.5 Cross-Modal Generation Coherence

Current world models often struggle to achieve consistent cross-modal generation, wherein visual, geometric, and semantic modalities must jointly interact to form a coherent representation of the environment. Misalignment can result in generated imagery that conflicts with the underlying 3D structure, undermining reliability in downstream perception and planning tasks. Overcoming these issues requires integrated architectures that jointly learn from diverse sensor data while enforcing strict consistency constraints during generation [77], [226]. Furthermore, ensuring fine-grained spatial alignment and temporal synchronization is crucial for accurately modeling the dynamic interactions in realistic driving environments. Future research should target this fundamental challenge to harmonize diverse data streams.

7 Conclusion

This survey has presented the first systematic review of **3D and 4D world modeling and generation**, clarifying definitions, organizing methods into a hierarchical taxonomy across VideoGen, OccGen, and LiDARGen, and summarizing datasets, evaluations, and applications. By shifting focus from purely visual realism to geometry-grounded modeling, native 3D and 4D representations enable models to achieve plausibility, controllability, and physical consistency, serving roles as data engines, action interpreters, neural simulators, and scene reconstructors. Despite rapid progress, challenges remain in scaling to real-world complexity, aligning multi-modal signals, and establishing standardized evaluation for controllability, safety, and generalization. Looking forward, unifying generative and predictive paradigms, integrating language and reasoning, and advancing simulation and digital twin ecosystems represent promising directions. Equally important will be community efforts in creating open benchmarks, reproducible codebases, and large-scale datasets tailored for 3D/4D world models, which can accelerate progress and ensure comparability across methods. We hope this survey provides both a coherent foundation and a forward-looking roadmap for advancing robust, interpretable, and generalizable world models to power the next generation of embodied AI.

TABLE 14: Summary of the evaluation metrics used for evaluating the quality of [1]**Generation**, [2]**Forecasting**, [3]**Planning**, [4]**Reconstruction**, and [5]**Downstream Tasks** for the **VideoGen**, **OccGen**, and **LiDARGen** models in 2D, 3D, and 4D tasks.

Abbr.	-	Full Name	Description	Ref.
⚖ **Generation Quality - Perceptual Fidelity**				
FID	↓	Fréchet Inception Distance	Statistical distance between multivariate Gaussians fitted to Inception features of real and generated samples, measuring distributional similarity.	[246]
FVD	↓	Fréchet Video Distance	Statistical distance between multivariate Gaussians fitted to I3D features [318] of real and generated videos, capturing temporal coherence.	[247]
FRD	↓	Fréchet Range Distance	Statistical distance between Gaussians fitted to RangeNet++ features [319] extracted from LiDAR range images, assessing distributional fidelity.	[206]
FPD	↓	Fréchet Point Cloud Distance	Statistical distance between Gaussians fitted to PointNet features [62] of raw 3D point clouds, evaluating geometric realism.	[248]
FSVD	↓	Fréchet Sparse Volume Distance	Statistical distance between Gaussians fitted to volumetric encoder features of sparse voxel inputs, capturing volumetric structure.	[209]
FPVD	↓	Fréchet Point Volume Distance	Statistical distance between Gaussians fitted to volumetric encoder features of hybrid point–voxel representations, measuring fidelity.	[209]
F3D	↓	Fréchet 3D Distance	Statistical distance between Gaussians fitted to occupancy grid features, evaluating volumetric realism in generated 3D data.	[176]
S-FRD	↓	Semantic Fréchet Range Distance	Class-aware extension of FRD that incorporates semantic labels for improved alignment of LiDAR range distributions.	[215]
S-FPD	↓	Semantic Fréchet Point Distance	Class-aware extension of FPD that integrates semantic labels to assess alignment of 3D point cloud distributions.	[215]
KID	↓	Kernel Inception Distance	Maximum Mean Discrepancy between Inception features using a polynomial kernel, providing an unbiased distributional similarity measure.	[320]
IS	↑	Inception Score	Evaluates image realism by rewarding confident and diverse class predictions from a pretrained Inception classifier, without real reference data.	[249]
IQ	↑	Image Quality	Predicts perceptual image quality by estimating human opinion scores with a learned quality assessor [321], without ground truth references.	[76]
⚖ **Generation Quality - Statistical Fidelity**				
PR	↑	Precision-Recall	Reports sample fidelity as precision and distributional coverage as recall, characterizing closeness to the real data manifold.	[18]
SWD	↓	Sliced Wasserstein Distance	Mean Wasserstein distance over multiple random 1D projections of image patches at different scales, reflecting distributional similarity.	[212]
JSD	↓	Jensen–Shannon Divergence	Symmetric divergence measuring similarity between occupancy histograms of real and generated scenes, lower indicating better alignment.	[205]
MMD	↓	Minimum Matching Distance	Average Chamfer distance from each real sample to its nearest generated neighbor, quantifying geometric fidelity.	[205]
COV	↑	Coverage	Fraction of real samples matched by at least one generated output, measuring generative diversity and recall.	[224]
1-NNA	-	1-Nearest-Neighbor Accuracy	Overlap test using a 1-NN classifier trained across sets, where accuracy near 50% indicates distributional equivalence.	[224]
Diversity	↑	-	Degree of variability across generated outputs for fixed prompts, often measured via pixel- or feature-wise variance.	[91]
⚖ **Generation Quality - Spatial Consistency**				
VCS	↑	View Consistency Score	Summation of LoFTR keypoint confidences [322] across overlapping views, evaluating multi-view geometric consistency and alignment quality.	[72]
KPM	↑	Key Points Matching	Ratio of successfully matched keypoints between adjacent generated and real views, reflecting geometric alignment quality.	[102]
DAS	↓	Depth Alignment Score	Statistical discrepancy between projected point clouds and estimated monocular depth [323], measuring scene-level depth consistency.	[226]
⚖ **Generation Quality - Temporal Consistency**				
CTC	↑	CLIP Temporal Consistency	Cosine similarity of CLIP features [250] across consecutive frames, measuring temporal stability and smoothness in generated video sequences.	[76]
DTC	↑	DINO Temporal Consistency	Cosine similarity of DINO features [251] across adjacent frames, evaluating temporal coherence and smooth transitions in generated sequences.	[76]

Note: Continued on next page

Abbr.	-	Full Name	Description	Ref.
TTCE	↓	Temporal Transformation Consistency Error	Registration error between temporally generated and ground-truth point clouds, evaluating motion alignment and temporal consistency.	[49]
CTC	↓	Chamfer Temporal Consistency	Chamfer distance between generated point clouds across different timestamps, quantifying temporal stability and geometric coherence.	[49]
ICP	↓	ICP Energy / Outlier	Registration residuals and outlier ratios from Iterative Closest Point alignment of LiDAR frames, detecting temporal jitter and misalignment.	[227]
⚖ **Generation Quality - Subject Consistency**				
SC	↑	Subject Consistency	Cosine similarity of subject-region features [251] across frames, evaluating identity persistence and stability in generated video sequences.	[252]
FDC	↑	Foreground Detection Confidence	Confidence scores of detected foreground objects in generated samples using a pretrained detector, reflecting semantic plausibility and realism.	[49]
CFCA	↑	Conditional Foreground Classification Accuracy	Semantic consistency of generated objects evaluated by classification accuracy with a pretrained object classifier, conditioned on ground truth.	[49]
CFSC	↑	Conditional Foreground Spatial Consistency	Mean IoU between 3D boxes regressed by a conditional VAE and ground truth boxes, assessing geometric alignment under conditioning.	[49]
⚖ **Generation Quality - Controllability**				
CDA	↑	Conditional Detection Accuracy	Standard detection accuracy from a pretrained 3D detector applied to generated point clouds with box conditioning, measuring semantic fidelity.	[49]
CLIP-Sim	↑	CLIP Similarity	Average cosine similarity between CLIP embeddings of generated and reference frames, reflecting semantic alignment across modalities.	[86]
MAE	↓	Mean Absolute Error	Difference in predicted versus reference object counts within scene graphs, assessing accuracy of graph-level controllability.	[193]
JI	↑	Jaccard Index	Overlap ratio of predicted and reference category sets within scene graphs, evaluating graph-level semantic consistency.	[193]
RotErr	↓	Rotation Error	Angular difference between recovered and target camera trajectories, quantifying rotational alignment error.	[129]
TransErr	↓	Translation Error	Euclidean distance between recovered and target camera trajectories, quantifying translational alignment error.	[129]
⚖ **Generation Quality - Human Preference**				
VQ	↑	Visual Quality (2AFC)	Win rates for perceptual visual quality in two-alternative forced choice comparisons, reflecting human-preferred realism of generations.	[131]
MR	↑	Motion Rationality (2AFC)	Win rates for perceived motion rationality in two-alternative forced choice settings, evaluating naturalness of temporal dynamics.	[131]
DMOS	↑	Differential Mean Opinion Score	Average human-rated alignment with conditioning constraints (e.g., scene graphs), providing a relative perceptual quality measure.	[193]
⚖ **Forecasting Quality - Spatial Predictive Accuracy**				
L1 Error	↓	Frame Mean Absolute Error	Pixel- or depth-space L1 distance between predicted and ground-truth frames, quantifying reconstruction fidelity and short-horizon accuracy.	[106]
L2 Error	↓	Frame Mean Squared Error	Pixel- or depth-space L2 distance between predicted and ground-truth frames, reflecting average squared reconstruction deviation.	[106]
IoU$_c$	↑	IoU at Current Timestamp	Intersection-over-Union between predicted and reference occupancy maps at the current frame, assessing immediate prediction quality.	[171]
IoU$_f$	↑	IoU at Future Timestamp	Intersection-over-Union between predicted and reference occupancy maps at a fixed future horizon, capturing long-range prediction quality.	[171]
IoU$_{wf}$	↑	IoU at Weighted Future Timestamp	Weighted average Intersection-over-Union across multiple future frames, emphasizing near-term predictions for smoother accuracy assessment.	[171]
CD	↓	Chamfer Distance	Bidirectional nearest-neighbor distance between point clouds from ray-cast predicted and ground-truth occupancy, measuring geometric fidelity.	[19]
L1 Med	↓	Median L1 Depth Error	Median absolute depth error along LiDAR rays after projection, robustly quantifying accuracy against outliers.	[19]
AbsRel Med	↓	Median Absolute Relative Error	Median of relative depth errors across all LiDAR rays, providing a scale-aware and robust measure of accuracy.	[19]
L1 Mean	↓	Mean L1 Depth Error	Mean absolute depth error along projected LiDAR rays, reflecting average deviation in meters from reference.	[19]

Note: Continued on next page

Abbr.	-	Full Name	Description	Ref.
AbsRel Mean	↓	Mean Absolute Relative Error	Mean of relative depth errors across all rays, capturing overall scale-consistent accuracy of depth predictions.	[19]
⚖ Forecasting Quality - Temporal Predictive Accuracy				
KODP	↑	Key Object Dimension Probability	Probability-based measure penalizing implausible object dimensions using category priors, encouraging physically realistic and consistent generation.	[194]
TFSC	↑	Temporal Foreground Shape Consistency	Voxel-level Intersection-over-Union of dynamic object instances across consecutive frames, ensuring shape persistence and temporal stability.	[194]
TBEC	↑	Temporal Background Environment Consistency	Consistency of static voxels under ego-motion compensation, validating environmental rigidity and long-term background stability.	[194]
⚖ Planning Quality - Open-Loop Planning				
ADE	↓	Average Displacement Error	Mean displacement error between predicted trajectories and expert waypoints across the horizon, reflecting overall trajectory accuracy.	[96]
FDE	↓	Final Displacement Error	Displacement error at the final predicted waypoint compared with expert trajectories, emphasizing long-term accuracy.	[96]
SLE	↓	Speed L1 Error	Mean absolute error of predicted speed control signals.	[101]
SALE	↓	Steer Angle L1 Error	Mean absolute error of predicted steering angle control signals.	[101]
CR	↓	Collision rate	Fraction of rollouts in which the controlled vehicle collides with surrounding agents or obstacles, indicating safety risk.	[102]
PDMS	↑	Predictive Driver Model Score	Aggregate score combining progress, spacing, and comfort after discarding unsafe rollouts, approximating human-like driving quality.	[99]
EPDMS	↑	Extended Predictive Driver Model Score	Extended version of PDMS that includes nine additional factors to reflect rule adherence and recovery behaviors.	[253]
AHE	↓	Average Heading Error	Mean absolute angular deviation between predicted and expert heading over the trajectory horizon, measuring orientation accuracy.	[98]
FHE	↓	Final Heading Error	Absolute angular deviation of predicted heading from expert at the final timestep, reflecting terminal orientation accuracy.	[98]
MR	↓	Miss Rate	Fraction of prediction timesteps where displacement error exceeds horizon-specific thresholds, reflecting failure in trajectory coverage.	[98]
⚖ Planning Quality - Closed-Loop Planning				
SR	↑	Success Rate	Percentage of navigation episodes that successfully reach the goal within a fixed time budget, indicating overall task completion.	[100]
ID	↑	Infraction Distance	Average driving distance between two infractions, with longer distances reflecting safer and more reliable policy behavior.	[100]
ADS	↑	Arena Driving Score	Composite score combining route completion metrics with PDMS to summarize closed-loop driving performance in Arena environments.	[119]
NAC	↑	No At-Fault Collisions	Fraction of scenarios without ego-fault collisions, focusing exclusively on responsibility-aware collision evaluation.	[98]
DAC	↑	Drivable-Area Compliance	Boolean evaluation that checks whether the ego vehicle remains inside drivable polygons throughout the rollout.	[98]
DDC	↑	Driving-Direction Compliance	Boolean evaluation verifying that ego motion remains aligned with the designated lane's legal driving direction.	[98]
MP	↑	Making Progress	Boolean check confirming that the ego vehicle makes sufficient forward route progress within the evaluation horizon.	[98]
TTC	↑	Time-to-Collision	Boolean verification that the time-to-collision value exceeds safety thresholds, preventing imminent crashes.	[98]
PAR	↑	Progress Along Route	Ratio of ego-vehicle progress compared to expert trajectory progress along the same route, reflecting efficiency.	[98]
SLC	↑	Speed-Limit Compliance	Score penalizing magnitude and duration of speed-limit violations, higher values indicating safer speed adherence.	[98]
Comfort	↑	Driving Comfort	Penalization of excessive jerk, acceleration, or yaw-rate, reflecting ride quality and passenger comfort.	[98]
⚖ Reconstruction Quality - Photometric Fidelity				
PSNR	↑	Peak Signal-to-Noise Ratio	Logarithmic ratio of maximum possible signal power to reconstruction error, measuring pixel-level fidelity of generated images	[258]

Note: Continued on next page

Abbr.	-	Full Name	Description	Ref.
SSIM	↑	Structural Similarity Index Measure	Quality index considering structural similarity, luminance consistency, and contrast preservation between generated and reference images.	[259]
LPIPS	↓	Learned Perceptual Image Patch Similarity	Feature-space distance between deep network activations of image patches, quantifying perceptual realism beyond pixel fidelity.	[260]
⚖ **Reconstruction Quality - View Changing Consistency**				
NTA-IoU	↑	Novel Trajectory Agent IoU	Intersection-over-Union between projected 3D bounding boxes of foreground agents and detected 2D boxes under novel viewpoints.	[50]
NTL-IoU	↑	Novel Trajectory Lane IoU	Intersection-over-Union between projected lane structures and detected lane markings from novel viewpoints, evaluating background alignment.	[50]
⚖ **Downstream Evaluation - Detection**				
mAP	↑	Mean Average Precision	Average precision computed over multiple IoU thresholds for 2D detection boxes on standard benchmarks, reflecting detection accuracy.	[261]
mAP-3D	↑	Mean Average Precision in 3D	Average precision for 3D bounding boxes, integrating precision-recall across multiple IoU thresholds in 3D space.	[10]
LET-3D-AP	↑	Average Precision with Longitudinal Error Tolerance	3D average precision using LET-IoU, allowing depth shifts along the camera ray within a tolerance margin.	[324]
LET-3D-APL	↑	Longitudinal Affinity Weighted LET-3D-AP	Weighted LET-3D-AP that penalizes larger longitudinal corrections, improving realism in depth-sensitive evaluation.	[324]
mATE	↓	Mean Average Translation Error	Mean L2 distance between predicted and ground-truth object centers for matched true positives, evaluating localization accuracy.	[10]
mASE	↓	Mean Average Scale Error	Mean scale discrepancy defined as one minus IoU for matched true positives, reflecting object size accuracy.	[10]
mAOE	↓	Mean Average Orientation Error	Mean absolute yaw error of matched true positives, quantifying accuracy of predicted object orientation.	[10]
mAVE	↓	Mean Average Velocity Error	Mean L2 difference between predicted and ground-truth object velocities for matched true positives, measuring motion accuracy.	[10]
mAAE	↓	Mean Average Attribute Error	Mean error in predicting semantic attributes for matched true positives, computed as one minus attribute accuracy.	[10]
NDS	↑	nuScenes Detection Score	Composite metric combining mAP with normalized true positive errors (mATE, mASE, mAOE, mAVE, mAAE), reflecting holistic detection quality.	[10]
⚖ **Downstream Evaluation - Segmentation**				
mIoU	↑	Mean Intersection over Union	Average Intersection-over-Union across all semantic classes in 2D images or 3D point clouds, measuring segmentation accuracy.	[325]
BEV-Map-IoU	↑	Bird's-Eye-View Map IoU	Class-wise Intersection-over-Union for freespace, lane, and dynamic agents in BEV compared with HD maps, evaluating scene consistency.	[326]
⚖ **Downstream Evaluation - Tracking**				
MOTA	↑	Multi-Object Tracking Accuracy	Composite metric aggregating false positives, false negatives, and identity switches into a single score, evaluating overall tracking reliability.	[262]
MOTP	↑	Multi-Object Tracking Precision	Mean Intersection-over-Union between matched predictions and ground truth across frames, quantifying spatial localization precision.	[262]
3D-AMOTA	↑	Average Multi-Object Tracking Accuracy in 3D	3D extension of MOTA averaged across recall thresholds, reducing sensitivity to threshold choices while evaluating tracking accuracy.	[327]
3D-AMOTP	↑	Average Multi-Object Tracking Precision in 3D	3D extension of MOTP averaging localization precision (IoU or center distance) over recall thresholds, reflecting robustness.	[327]
⚖ **Downstream Evaluation - Occupancy Prediction**				
Occupancy-IoU	↑	Occupancy Intersection over Union	Intersection-over-Union between predicted and labeled voxel occupancies at class- or scene-level granularity, reflecting occupancy accuracy.	[328]
VPQ	↑	Voxelized Panoptic Quality	Panoptic quality metric for voxelized outputs, combining semantic segmentation accuracy with instance detection recall into a unified score.	[171]
⚖ **Downstream Evaluation - VQA**				
Top-1 Acc	↑	Visual Question Answering Top-1 Accuracy	Exact-match accuracy of predicted answers across diverse question categories in autonomous driving VQA tasks, measuring reasoning reliability.	[181]

REFERENCES

[1] J. Bruce, M. D. Dennis, A. Edwards *et al.*, "Genie: Generative interactive environments," in *Int. Conf. Learn. Represent.*, 2024.

[2] A. Bardes *et al.*, "Revisiting feature prediction for learning visual representations from video," *arXiv preprint arXiv:2404.08471*, 2024.

[3] J. Peper *et al.*, "Four principles for physically interpretable world models," *arXiv preprint arXiv:2503.02143*, 2025.

[4] J. Parker-Holder *et al.*, "Genie 2: A large-scale foundation world model," https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model, 2024.

[5] J. Ding, Y. Zhang, Y. Shang *et al.*, "Understanding world or predicting future? a comprehensive survey of world models," *ACM Computing Surveys*, 2024.

[6] J. Cho, F. D. Puspitasari, S. Zheng, J. Zheng, L.-H. Lee, T.-H. Kim, C. S. Hong, and C. Zhang, "Sora as an AGI world model? a complete survey on text-to-video generation," *arXiv preprint arXiv:2403.05131*, 2024.

[7] X. Mai *et al.*, "From efficient multimodal models to world models: A survey," *arXiv preprint arXiv:2407.00118*, 2024.

[8] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4195–4205.

[9] N. Silberman *et al.*, "Indoor segmentation and support inference from RGBD images," in *Eur. Conf. Comput. Vis.* Springer, 2012, pp. 746–760.

[10] H. Caesar, V. Bankiti, A. H. Lang *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 11 621–11 631.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3354–3361.

[12] L. Mescheder, M. Oechsle, M. Niemeyer *et al.*, "Occupancy networks: Learning 3D reconstruction in function space," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4460–4470.

[13] A.-Q. Cao and R. D. Charette, "MonoScene: Monocular 3D semantic scene completion," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 3991–4001.

[14] X. Tian, T. Jiang, L. Yun *et al.*, "Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 64 318–64 330.

[15] W. K. Fong, R. Mohan, J. V. Hurtado *et al.*, "Panoptic nuScenes: A large-scale benchmark for LiDAR panoptic segmentation and tracking," *IEEE Robot. Autom. Lett.*, vol. 7, pp. 3795–3802, 2022.

[16] J. Behley, M. Garbade, A. Milioto *et al.*, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9297–9307.

[17] L. Kong *et al.*, "LaserMix for semi-supervised LiDAR semantic segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 21 705–21 715.

[18] H. Bian *et al.*, "DynamicCity: Large-scale 4D occupancy generation from dynamic scenes," in *Int. Conf. Learn. Represent.*, 2025.

[19] L. Zhang *et al.*, "Copilot4D: Learning unsupervised world models for autonomous driving via discrete diffusion," in *Int. Conf. Learn. Represent.*, 2024.

[20] R. Gao *et al.*, "MagicDrive: Street view generation with diverse 3D geometry control," in *Int. Conf. Learn. Represent.*, 2023.

[21] R. Gao, K. Chen, Z. Li *et al.*, "MagicDrive3D: Controllable 3D generation for any-view rendering in street scenes," *arXiv preprint arXiv:2405.14475*, 2024.

[22] Y. Liu *et al.*, "La La LiDAR: Large-scale layout generation from LiDAR data," *arXiv preprint arXiv:2508.03691*, 2025.

[23] L. Kong, X. Xu, J. Ren *et al.*, "Multi-modal data-efficient 3D scene understanding for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3748–3765, 2025.

[24] Y. Li *et al.*, "Is your LiDAR placement optimized for 3D scene understanding?" in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 34 980–35 017.

[25] S. Zhou *et al.*, "RoboDreamer: Learning compositional world models for robot imagination," *arXiv preprint arXiv:2404.12377*, 2024.

[26] X. Zhou *et al.*, "Genesis: A generative and universal physics engine for robotics and beyond," *arXiv preprint arXiv:2401.01454*, 2024.

[27] A. Fu *et al.*, "Exploring the interplay between video generation and world models in autonomous driving: A survey," *arXiv preprint arXiv:2411.02914*, 2024.

[28] P. Fung, Y. Bachrach, A. Celikyilmaz, K. Chaudhuri, D. Chen, W. Chung, E. Dupoux, H. Gong, H. Jégou, A. Lazaric *et al.*, "Embodied AI agents: Modeling the world," *arXiv preprint arXiv:2506.22355*, 2025.

[29] M. Yu *et al.*, "TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models," *arXiv preprint arXiv:2503.05638*, 2025.

[30] H. Team *et al.*, "HunyuanWorld 1.0: Generating immersive, explorable, and interactive 3D worlds from words or pixels," *arXiv preprint arXiv:2507.21809*, 2025.

[31] T. Huang *et al.*, "Voyager: Long-range and world-consistent video diffusion for explorable 3D scene generation," *arXiv preprint arXiv:2506.04225*, 2025.

[32] X. Yang *et al.*, "Hunyuan3D 1.0: A unified framework for text-to-3D and image-to-3D generation," *arXiv preprint arXiv:2411.02293*, 2024.

[33] Z. Zhao *et al.*, "Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation," *arXiv preprint arXiv:2501.12202*, 2025.

[34] Z. Lai *et al.*, "Hunyuan3D 2.5: Towards high-fidelity 3D assets generation with ultimate details," *arXiv preprint arXiv:2506.16504*, 2025.

[35] X. Ren, Y. Lu, T. Cao *et al.*, "Cosmos-Drive-Dreams: Scalable synthetic driving data generation with world foundation models," *arXiv preprint arXiv:2506.09042*, 2025.

[36] A. Azzolini *et al.*, "Cosmos-Reason1: From physical common sense to embodied reasoning," *arXiv preprint arXiv:2503.15558*, 2025.

[37] N. Agarwal *et al.*, "Cosmos world foundation model platform for physical AI," *arXiv preprint arXiv:2501.03575*, 2025.

[38] H. A. Alhaija *et al.*, "Cosmos-Transfer1: Conditional world generation with adaptive multimodal control," *arXiv preprint arXiv:2503.14492*, 2025.

[39] P. J. Ball, J. Bauer, F. Belletti *et al.*, "Genie 3: A new frontier for world models," https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, 2025.

[40] M. Assran, A. Bardes, D. Fan *et al.*, "V-JEPA 2: Self-supervised video models enable understanding, prediction and planning," *arXiv preprint arXiv:2506.09985*, 2025.

[41] J. Ding *et al.*, "Understanding world or predicting future? a comprehensive survey of world models," *arXiv preprint arXiv:2411.14499*, 2024.

[42] H. Xu, J. Chen, S. Meng *et al.*, "A survey on occupancy perception for autonomous driving: The information fusion perspective," *Information Fusion*, vol. 114, p. 102671, 2025.

[43] X. Long *et al.*, "A survey: Learning embodied intelligence from physical simulators and world models," *arXiv preprint arXiv:2507.00917*, 2025.

[44] Y. Guan, H. Liao, Z. Li *et al.*, "World models for autonomous driving: An initial survey," *IEEE Trans. Intell. Veh.*, pp. 1–17, 2024.

[45] X. Yan *et al.*, "Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities," *arXiv preprint arXiv:2401.08045*, 2024.

[46] S. Tu *et al.*, "The role of world models in shaping autonomous driving: A comprehensive survey," *arXiv preprint arXiv:2502.10498*, 2025.

[47] B. Kang *et al.*, "How far is video generation from world model: A physical law perspective," in *Int. Conf. Mach. Learn.* PMLR, 2025.

[48] Z. Zhu *et al.*, "Is Sora a world simulator? a comprehensive survey on general world models and beyond," *arXiv preprint arXiv:2405.03520*, 2024.

[49] A. Liang *et al.*, "LiDARCrafter: Dynamic 4D world modeling from LiDAR sequences," *arXiv preprint arXiv:2508.03692*, 2025.

[50] G. Zhao, C. Ni, X. Wang *et al.*, "DriveDreamer4D: World models are effective data machines for 4D driving scene representation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025.

[51] M. Lin *et al.*, "Exploring the evolution of physics cognition in video generation: A survey," *arXiv preprint arXiv:2503.21765*, 2025.

[52] Y. Wang *et al.*, "Generative ai for autonomous driving: Frontiers and opportunities," *arXiv preprint arXiv:2505.08854*, 2025.

[53] C. Zhao *et al.*, "Edge general intelligence through world models and agentic AI: Fundamentals, solutions, and challenges," *arXiv preprint arXiv:2508.09561*, 2025.

[54] N. Xie *et al.*, "From 2D to 3D cognition: A brief survey of general world models," *arXiv preprint arXiv:2506.20134*, 2025.

[55] B. Wen *et al.*, "3d scene generation: A survey," *arXiv preprint arXiv:2505.05474*, 2025.

[56] T. Feng, W. Wang, and Y. Yang, "A survey of world models for autonomous driving," *arXiv preprint arXiv:2501.11260*, 2025.

[57] C. Zhao *et al.*, "World models for cognitive agents: Transforming edge intelligence in future networks," *arXiv preprint arXiv:2506.00417*, 2025.

[58] X. Wang, Z. Zhu, G. Huang *et al.*, "DriveDreamer: Towards real-world-drive world models for autonomous driving," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 55–72.

[59] K. Zhang *et al.*, "Epona: Autoregressive diffusion world model for autonomous driving," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[60] A. Hu *et al.*, "GAIA-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[61] S. Peng *et al.*, "Convolutional occupancy networks," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 523–540.

[62] C. R. Qi *et al.*, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 652–660.

[63] T. Khurana *et al.*, "Point cloud forecasting as a proxy for 4D occupancy forecasting," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 1116–1124.

[64] X. Xu *et al.*, "4D contrastive superflows are dense 3D representation learners," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 58–80.

[65] L. Kong, Y. Liu, X. Li *et al.*, "Robo3D: Towards robust and reliable 3D perception against corruptions," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19 994–20 006.

[66] D. P. Kingma, M. Welling *et al.*, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[67] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Comm. of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[68] J. Ho *et al.*, "Denoising diffusion probabilistic models," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[69] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[70] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[71] K. Tian *et al.*, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 84 839–84 865.

[72] A. Swerdlow, R. Xu, and B. Zhou, "Street-view image generation from a bird's-eye view layout," *IEEE Robot. Autom. Lett.*, vol. 9, no. 4, pp. 3578–3585, 2024.

[73] K. Yang, E. Ma, J. Peng *et al.*, "BEVControl: Accurately controlling street-view elements with multi-perspective consistency via BEV sketch layout," *arXiv preprint arXiv:2308.01661*, 2023.

[74] R. Gao, K. Chen, B. Xiao *et al.*, "MagicDrive-V2: High-resolution long video generation for autonomous driving with adaptive control," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[75] L. Li, W. Qiu, Y. Cai *et al.*, "SyntheOcc: Synthesize geometric-controlled street view images through 3D semantic mpis," *arXiv preprint arXiv:2410.00337*, 2024.

[76] J. Jiang *et al.*, "DiVE: DiT-based video generation with enhanced control," *arXiv preprint arXiv:2409.01595*, 2024.

[77] B. Li, J. Guo, H. Liu *et al.*, "UniScene: Unified occupancy-centric driving scene generation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 11 971–11 981.

[78] H. Dong, X. Wang, D. Lin *et al.*, "NoiseController: Towards consistent multi-view video generation via noise decomposition and collaboration," *arXiv preprint arXiv:2504.18448*, 2025.

[79] J. Zhang, H. Sheng, S. Cai *et al.*, "PerLDiff: Controllable street view synthesis using perspective-layout diffusion models," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[80] Y. Wen, Y. Zhao, Y. Liu *et al.*, "Panacea: Panoramic and controllable video generation for autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 6902–6912.

[81] X. Li *et al.*, "DrivingDiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 469–485.

[82] B. Huang, Y. Wen, Y. Zhao *et al.*, "SubjectDrive: Scaling generative data in autonomous driving via subject control," in *AAAI Conf. Artifi. Intell.*, vol. 39, 2025, pp. 3617–3625.

[83] B. Xie, Y. Liu, T. Wang *et al.*, "Glad: A streaming scene generator for autonomous driving," in *Int. Conf. Learn. Represent.*, 2025.

[84] J. Wang, Y. Yao, X. Feng *et al.*, "STAGE: A stream-centric generative world model for long-horizon driving-scene simulation," *arXiv preprint arXiv:2506.13138*, 2025.

[85] Y. Zhang, S. Gong, K. Xiong *et al.*, "BEVWorld: A multimodal world model for autonomous driving via unified BEV latent space," *arXiv preprint arXiv:2407.05679*, 2024.

[86] E. Ma, L. Zhou, T. Tang *et al.*, "Unleashing generalization of end-to-end autonomous driving with controllable long video generation," *arXiv preprint arXiv:2406.01349*, 2024.

[87] G. Zhao, X. Wang, Z. Zhu *et al.*, "DriveDreamer-2: LLM-enhanced world models for diverse driving video generation," in *AAAI Conf. Artifi. Intell.*, vol. 39, 2025, pp. 10 412–10 420.

[88] Y. Zhou, N. Ye, W. Ljungbergh *et al.*, "Decoupled diffusion sparks adaptive scene generation," *arXiv preprint arXiv:2504.10485*, 2025.

[89] Z. Xu, B. Li, H. Gao *et al.*, "Challenger: Affordable adversarial driving video generation," *arXiv preprint arXiv:2505.15880*, 2025.

[90] J. Lu, Z. Huang, Z. Yang *et al.*, "WoVoGen: World volume-aware diffusion for controllable multi-camera driving scene generation," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 329–345.

[91] Y. Zhou, M. Simon, Z. M. Peng *et al.*, "SimGen: Simulator-conditioned driving scene generation," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 48 838–48 874.

[92] R. Chen *et al.*, "UniMLVG: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving," *arXiv preprint arXiv:2412.04842*, 2024.

[93] Z. Yang, X. Guo, C. Ding *et al.*, "Physical informed driving world model," *arXiv preprint arXiv:2412.08410*, 2024.

[94] A. Chen, W. Zheng, Y. Wang *et al.*, "GeoDrive: 3D geometry-informed driving world model with precise action control," *arXiv preprint arXiv:2505.22421*, 2025.

[95] P. Sun, H. Kretzschmar, X. Dotiwalla *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 2446–2454.

[96] J. Yang, S. Gao, Y. Qiu *et al.*, "Generalized predictive model for autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 662–14 672.

[97] B. Wilson, W. Qi, T. Agarwal *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[98] H. Caesar, J. Kabzan, K. S. Tan *et al.*, "nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.

[99] D. Dauner, M. Hallgarten, T. Li *et al.*, "NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 28 706–28 719.

[100] A. Dosovitskiy *et al.*, "CARLA: An open urban driving simulator," in *Conf. Robot Learn.* PMLR, 2017, pp. 1–16.

[101] F. Jia, W. Mao, Y. Liu *et al.*, "ADriver-I: A general world model for autonomous driving," *arXiv preprint arXiv:2311.13549*, 2023.

[102] Y. Wang *et al.*, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 749–14 759.

[103] S. Gao, J. Yang, L. Chen *et al.*, "Vista: A generalizable driving world model with high fidelity and versatile controllability," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 91 560–91 596.

[104] Y. Wen, Y. Zhao, Y. Liu *et al.*, "Panacea+: Panoramic and controllable video generation for autonomous driving," *arXiv preprint arXiv:2408.07605*, 2024.

[105] J. Mei, T. Hu, X. Yang *et al.*, "DreamForge: Motion-aware autoregressive video generation for multi-view driving scenes," *arXiv preprint arXiv:2409.04003*, 2024.

[106] Z. Wu, J. Ni, X. Wang *et al.*, "HoloDrive: Holistic 2D-3D multi-modal street scene generation for autonomous driving," *arXiv preprint arXiv:2412.01407*, 2024.

[107] X. Guo *et al.*, "InfinityDrive: Breaking time limits in driving world models," *arXiv preprint arXiv:2412.01522*, 2024.

[108] H. Lu, X. Wu, S. Wang *et al.*, "Seeing beyond views: Multi-view driving scene video generation with holistic attention," *arXiv preprint arXiv:2412.03520*, 2024.

[109] W. Zheng *et al.*, "Doe-1: Closed-loop autonomous driving with large world model," *arXiv preprint arXiv:2412.09627*, 2024.

[110] B. Li *et al.*, "OccScene: Semantic occupancy-based cross-task mutual learning for 3D scene generation," *arXiv preprint arXiv:2412.11183*, 2024.

[111] Y. Chen, Y. Wang, and Z. Zhang, "DrivingGPT: Unifying driving world modeling and planning with multi-modal autoregressive transformers," *arXiv preprint arXiv:2412.18607*, 2024.

[112] X. Hu, W. Yin, M. Jia *et al.*, "DrivingWorld: Constructing world model for autonomous driving via video GPT," *arXiv preprint arXiv:2412.19505*, 2024.

[113] H. Li, Z. Yang, Z. Qian *et al.*, "DualDiff: Dual-branch diffusion model for autonomous driving with semantic fusion," in *IEEE Int. Conf. Robot. Autom.*, 2025.

[114] W. Wu, X. Guo, W. Tang *et al.*, "DriveScape: Towards high-resolution controllable multi-view driving video generation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 17 187–17 196.

[115] T. Yan, D. Wu, W. Han *et al.*, "DrivingSphere: Building a high-fidelity 4D world for closed-loop simulation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 27 531–27 541.

[116] M. Hassan, S. Stapf, A. Rahimi *et al.*, "GEM: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 22 404–22 415.

[117] J. Ni, Y. Guo, Y. Liu *et al.*, "MaskGWM: A generalizable driving world model with video mask reconstruction," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 22 381–22 391.

[118] Y. Wu, H. Zhang, T. Lin *et al.*, "Generating multimodal driving scenes via next-scene prediction," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 6844–6853.

[119] X. Yang, L. Wen, Y. Ma *et al.*, "DriveArena: A closed-loop generative simulation platform for autonomous driving," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[120] Y. Lu, X. Ren, J. Yang *et al.*, "InfiniCube: Unbounded and controllable dynamic 3D driving scene generation with world-guided video models," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[121] J. Guo, Y. Ding, X. Chen *et al.*, "DiST-4D: Disentangled spatiotemporal diffusion with metric depth for 4D driving scene generation," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[122] F. Bartoccioni, E. Ramzi, V. Besnier *et al.*, "VaViM and VaVAM: Autonomous driving through video generative modeling," *arXiv preprint arXiv:2502.15672*, 2025.

[123] Z. Yang, Z. Qian, X. Li *et al.*, "DualDiff+: Dual-branch diffusion for high-fidelity video generation with reward guidance," *arXiv preprint arXiv:2503.03689*, 2025.

[124] D. Liang, D. Zhang, X. Zhou *et al.*, "Seeing the future, perceiving the future: A unified driving world model for future generation and perception," *arXiv preprint arXiv:2503.13587*, 2025.

[125] H. Wang, D. Liu, H. Xie *et al.*, "MiLA: Multi-view intensive-fidelity long-term video generation world model for autonomous driving," *arXiv preprint arXiv:2503.15875*, 2025.

[126] L. Russell, A. Hu, L. Bertoni *et al.*, "GAIA-2: A controllable multi-view generative world model for autonomous driving," *arXiv preprint arXiv:2503.20523*, 2025.

[127] Y. Ji, Z. Zhu, Z. Zhu *et al.*, "CoGen: 3D consistent video generation via adaptive conditioning for autonomous driving," *arXiv preprint arXiv:2503.22231*, 2025.

[128] X. Li, C. Wu, Z. Yang *et al.*, "DriVerse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment," *arXiv preprint arXiv:2504.18576*, 2025.

[129] B. Jin, W. Li, B. Yang *et al.*, "PosePilot: Steering camera pose for generative world models with self-supervised depth," *arXiv preprint arXiv:2505.01729*, 2025.

[130] X. Wang and P. Peng, "ProphetDWM: A driving world model for rolling out future actions and videos," *arXiv preprint arXiv:2505.18650*, 2025.

[131] X. Wang, Z. Wu, and P. Peng, "LongDWM: Cross-granularity distillation for building a long-term driving world model," *arXiv preprint arXiv:2506.01546*, 2025.

[132] H. Wang, X. Ye, F. Tao *et al.*, "AdaWM: Adaptive world model-based planning for autonomous driving," *arXiv preprint arXiv:2501.13072*, 2025.

[133] P. Li and D. Cui, "Navigation-guided sparse scene representation for end-to-end autonomous driving," *arXiv preprint arXiv:2409.18341*, 2024.

[134] Y. Zheng, P. Yang, Z. Xing *et al.*, "World4Drive: End-to-end autonomous driving via intention-aware physical latent world model," *arXiv preprint arXiv:2507.00603*, 2025.

[135] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 10 684–10 695.

[136] A. Blattmann, T. Dockhorn, S. Kulal *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[137] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Comm. of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[138] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D gaussian splatting for real-time radiance field rendering," *ACM Trans. Graphics*, vol. 42, no. 4, 2023.

[139] T. Fischer, J. Kulhanek, S. R. Bulò *et al.*, "Dynamic 3D Gaussian fields for urban areas," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 80 466–80 494.

[140] N. Huang *et al.*, "S3Gaussian: Self-supervised street Gaussians for autonomous driving," *arXiv preprint arXiv:2405.20323*, 2024.

[141] H. Li, J. Li, D. Zhang *et al.*, "VDG: Vision-only dynamic Gaussian for driving simulation," *arXiv preprint arXiv:2406.18198*, 2024.

[142] Y. Ren *et al.*, "UniGaussian: Driving scene reconstruction from multiple camera models via unified Gaussian representations," *arXiv preprint arXiv:2411.15355*, 2024.

[143] Z. Yu, H. Wang, J. Yang *et al.*, "SGD: Street view synthesis with Gaussian splatting and diffusion prior," in *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2025, pp. 3812–3822.

[144] C. Peng, C. Zhang, Y. Wang *et al.*, "DeSiRe-GS: 4D street Gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 6782–6791.

[145] L. Fan, H. Zhang, Q. Wang *et al.*, "FreeSim: Toward free-viewpoint camera simulation in driving scenes," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 12 004–12 014.

[146] J. Zhou *et al.*, "FlexDrive: Toward trajectory flexibility in driving scene reconstruction and rendering," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025.

[147] Y. Chen, J. Zhang, Z. Xie *et al.*, "S-NeRF++: Autonomous driving simulation via neural reconstruction and generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 6, pp. 4358–4376, 2025.

[148] J. Mao, B. Li, B. Ivanovic *et al.*, "DreamDrive: Generative 4D scene modeling from street view images," *arXiv preprint arXiv:2501.00601*, 2025.

[149] Y. Zou *et al.*, "MuDG: Taming multi-modal diffusion with Gaussian splatting for urban scene reconstruction," *arXiv preprint arXiv:2503.10604*, 2025.

[150] J. Ge *et al.*, "Unraveling the effects of synthetic data on end-to-end autonomous driving," *arXiv preprint arXiv:2503.18108*, 2025.

[151] J. Jiang *et al.*, "RealEngine: Simulating autonomous driving in realistic context," *arXiv preprint arXiv:2505.16902*, 2025.

[152] X. Zhang *et al.*, "AccidentSim: Generating physically realistic vehicle collision videos from real-world accident reports," *arXiv preprint arXiv:2503.20654*, 2025.

[153] S. Mo *et al.*, "Dreamland: Controllable world creation with simulator and generative models," *arXiv preprint arXiv:2506.08006*, 2025.

[154] Y. Yan, H. Lin, C. Zhou *et al.*, "Street Gaussians: Modeling dynamic urban scenes with gaussian splatting," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 156–173.

[155] H. Zhou *et al.*, "HUGSIM: A real-time, photo-realistic and closed-loop simulator for autonomous driving," *arXiv preprint arXiv:2412.01718*, 2024.

[156] Z. Yuan *et al.*, "Uni-Gaussians: Unifying camera and LiDAR simulation with Gaussians for dynamic driving scenarios," *arXiv preprint arXiv:2503.08317*, 2025.

[157] Z. Chen, J. Yang, J. Huang *et al.*, "OmniRe: Omni urban scene reconstruction," in *Int. Conf. Learn. Represent.*, 2025.

[158] H. Lu *et al.*, "DrivingRecon: Large 4D Gaussian reconstruction model for autonomous driving," *arXiv preprint arXiv:2412.09043*, 2024.

[159] C. Ni, G. Zhao, X. Wang *et al.*, "ReconDreamer: Crafting world models for driving scene reconstruction via online restoration," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 1559–1569.

[160] L. Wang *et al.*, "Stag-1: Towards realistic 4D driving simulation with video generation model," *arXiv preprint arXiv:2412.05280*, 2024.

[161] G. Zhao *et al.*, "ReconDreamer++: Harmonizing generative and reconstructive models for driving scene representation," *arXiv preprint arXiv:2503.18438*, 2025.

[162] Y. Yan, Z. Xu, H. Lin *et al.*, "StreetCrafter: Street view synthesis with controllable video diffusion models," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 822–832.

[163] J. Wilson, J. Song, Y. Fu *et al.*, "MotionSC: Data set and network for real-time semantic mapping in dynamic environments," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 8439–8446, 2022.

[164] J. Houston, G. Zuidhof, L. Bergamini *et al.*, "One thousand and one hours: Self-driving motion prediction dataset," in *Conf. Robot Learn.* PMLR, 2021, pp. 409–418.

[165] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3292–3310, 2022.

[166] R. Xu *et al.*, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *IEEE Int. Conf. Robot. Autom.*, 2022, pp. 2583–2589.

[167] T. Khurana, P. Hu, A. Dave *et al.*, "Differentiable raycasting for self-supervised occupancy forecasting," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 353–369.

[168] J. Lee, W. Im, S. Lee, and S.-E. Yoong, "Diffusion probabilistic models for scene-scale 3D categorical data," *arXiv preprint arXiv:2301.00527*, 2023.

[169] C. Min *et al.*, "UniWorld: Autonomous driving pre-training via world models," *arXiv preprint arXiv:2308.07234*, 2023.

[170] C. Min, L. Xiao, D. Zhao *et al.*, "Multi-camera unified pre-training via 3D scene reconstruction," *IEEE Robot. Autom. Lett.*, vol. 9, no. 4, pp. 3243–3250, 2024.

[171] J. Ma *et al.*, "Cam4DOcc: Benchmark for camera-only 4D occupancy forecasting in autonomous driving applications," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 21 486–21 495.

[172] X. Ren, J. Huang, X. Zeng *et al.*, "XCube: Large-scale 3D generative modeling using sparse voxel hierarchies," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 4209–4219.

[173] J. Lee, S. Lee, C. Jo *et al.*, "SemCity: Semantic scene generation with triplane diffusion," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 28 337–28 347.

[174] C. Min, D. Zhao, L. Xiao *et al.*, "DriveWorld: 4D pre-trained scene understanding via world models for autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 15 522–15 533.

[175] B. Agro, Q. Sykora, S. Casas *et al.*, "UnO: Unsupervised occupancy fields for perception and forecasting," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 487–14 496.

[176] Y. Liu, X. Li, X. Li *et al.*, "Pyramid diffusion for fine 3D large scene generation," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 71–87.

[177] W. Zheng, W. Chen, Y. Huang *et al.*, "OccWorld: Learning a 3D occupancy world model for autonomous driving," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 55–72.

[178] J. Zhang, Q. Zhang, L. Zhang *et al.*, "Urban scene diffusion through semantic occupancy map," *arXiv preprint arXiv:2403.11697*, 2024.

[179] L. Wang, W. Zheng, Y. Ren *et al.*, "OccSora: 4D occupancy generation models as world simulators for autonomous driving," *arXiv preprint arXiv:2405.20337*, 2024.

[180] B. Lange and otherss, "Self-supervised multi-future occupancy forecasting for autonomous driving," *arXiv preprint arXiv:2407.21126*, 2024.

[181] J. Wei, S. Yuan, P. Li *et al.*, "OccLLaMA: An occupancy-language-action generative world model for autonomous driving," *arXiv preprint arXiv:2409.03272*, 2024.

[182] E. Guo *et al.*, "FSF-Net: Enhance 4D occupancy forecasting with coarse BEV scene flow for autonomous driving," *arXiv preprint arXiv:2409.15841*, 2024.

[183] S. Gu, W. Yin, B. Jin *et al.*, "DOME: Taming diffusion model into high-fidelity controllable occupancy world model," *arXiv preprint arXiv:2410.10429*, 2024.

[184] W. Zheng *et al.*, "GaussianAD: Gaussian-centric end-to-end autonomous driving," *arXiv preprint arXiv:2412.10371*, 2024.

[185] H. Zhang, Y. Xue, X. Yan *et al.*, "An efficient occupancy world model via decoupled dynamic flow and image-assisted training," *arXiv preprint arXiv:2412.13772*, 2024.

[186] Y. Yang, J. Mei, Y. Ma *et al.*, "Driving in the occupancy world: Vision-centric 4D occupancy forecasting and planning via world models for autonomous driving," in *AAAI Conf. Artifi. Intell.*, vol. 39, 2025, pp. 9327–9335.

[187] X. Li *et al.*, "Semi-supervised vision-centric 3D occupancy world model for autonomous driving," *arXiv preprint arXiv:2502.07309*, 2025.

[188] J. Chen *et al.*, "OccProphet: Pushing efficiency frontier of camera-only 4D occupancy forecasting with observer-forecaster-refiner framework," *arXiv preprint arXiv:2502.15180*, 2025.

[189] Z. Yan, W. Dong, Y. Shao *et al.*, "RenderWorld: World model with self-supervised 3D label," *arXiv preprint arXiv:2409.11356*, 2024.

[190] T. Xu, H. Lu, X. Yan *et al.*, "Occ-LLM: Enhancing autonomous driving with occupancy-based large language models," in *IEEE Int. Conf. Robot. Autom.*, 2025.

[191] J. Xu, X. Chen, J. Ma *et al.*, "Spatiotemporal decoupling for efficient vision-based occupancy forecasting," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 22 338–22 347.

[192] C. Diehl, Q. Sykora, B. Agro *et al.*, "DIO: Decomposable implicit 4D occupancy-flow world model," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 27 456–27 466.

[193] Y. Liu *et al.*, "Controllable 3D outdoor scene generation via scene graphs," *arXiv preprint arXiv:2503.07152*, 2025.

[194] Y. Wang *et al.*, "UniOcc: A unified benchmark for occupancy forecasting and prediction in autonomous driving," *arXiv preprint arXiv:2503.24381*, 2025.

[195] Z. Liao, P. Wei, R. Zhang *et al.*, "I2-World: Intra-inter tokenization for efficient dynamic 4D scene forecasting," *arXiv preprint arXiv:2507.09144*, 2025.

[196] H. Xu *et al.*, "Temporal triplane transformers as occupancy world models," *arXiv preprint arXiv:2503.07338*, 2025.

[197] Y. Shi, K. Jiang, Q. Meng *et al.*, "COME: Adding scene-centric forecasting control to occupancy world model," *arXiv preprint arXiv:2506.13260*, 2025.

[198] Y. Yang, A. Liang, J. Mei *et al.*, "X-Scene: Large-scale driving scene generation with high fidelity and flexible controllability," *arXiv preprint arXiv:2506.13558*, 2025.

[199] C. O. Tze, D. Dauner, Y. Liao *et al.*, "PrITTI: Primitive-based generation of controllable and editable 3D semantic scenes," *arXiv preprint arXiv:2506.19117*, 2025.

[200] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 17 981–17 993, 2021.

[201] L. Kong, Y. Liu, R. Chen *et al.*, "Rethinking range view representation for lidar segmentation," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 228–240.

[202] A. Liang, L. Kong, D. Lu *et al.*, "Perspective-invariant 3d object detection," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[203] K. Nakashima and R. Kurazumeg, "Learning to drop points for LiDAR scan synthesis," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 222–229.

[204] K. Nakashima, Y. Iwashita, and R. Kurazume, "Generative range imaging for learning scene priors of 3D LiDAR data," in *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 1256–1266.

[205] V. Zyrianov *et al.*, "Learning to generate realistic LiDAR point clouds," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 17–35.

[206] K. Nakashima and R. Kurazume, "LiDAR data synthesis with denoising diffusion probabilistic models," in *IEEE Int. Conf. Robot. Autom.*, 2024, pp. 14 724–1473.

[207] Y. Lipman *et al.*, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[208] K. Nakashima *et al.*, "Fast LiDAR data generation with rectified flows," in *IEEE Int. Conf. Robot. Autom.*, 2025.

[209] H. Ran, V. Guizilini, and Y. Wang, "Towards realistic scene generation with LiDAR diffusion models," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 738–14 748.

[210] Q. Hu *et al.*, "RangeLDM: Fast realistic LiDAR point cloud generation," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 115–135.

[211] L. Nunes *et al.*, "Towards generating realistic 3D semantic training data for autonomous driving," *arXiv preprint arXiv:2503.21449*, 2025.

[212] H. Haghighi *et al.*, "Taming transformers for realistic LiDAR point cloud generation," *arXiv preprint arXiv:2404.05505*, 2024.

[213] A. Van Den Oord *et al.*, "Neural discrete representation learning," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6309–6318.

[214] R. Faulkner *et al.*, "Simultaneous diffusion sampling for conditional LiDAR generation," *arXiv preprint arXiv:2410.11628*, 2024.

[215] D. Zhu, Y. Hu, Y. Liu *et al.*, "SPIRAL: Semantic-aware progressive LiDAR scene generation," *arXiv preprint arXiv:2505.22643*, 2025.

[216] Y. Liu, L. Kong, W. Yang *et al.*, "Veila: Panoramic LiDAR generation from a monocular RGB image," *arXiv preprint arXiv:2508.03690*, 2025.

[217] P. Xiao, Z. Shao, S. Hao *et al.*, "PandaSet: Advanced sensor suite dataset for autonomous driving," in *IEEE Int. Conf. Intell. Transport. Syst.*, 2021, pp. 3095–3101.

[218] M. Bijelic *et al.*, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 11 682–11 692.

[219] S. Wang, Z. Yu, X. Jiang *et al.*, "OmniDrive: A holistic LLM-agent framework for autonomous driving with 3D perception, reasoning, and planning," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 22 442–22 452.

[220] Y. Xiong *et al.*, "UltraLiDAR: Learning compact representations for LiDAR completion and generation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 1074–1083.

[221] Z. Yang *et al.*, "Visual point cloud forecasting enables scalable autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 673–14 684.

[222] L. Nunes *et al.*, "Scaling diffusion models to real-world 3D LiDAR scene completion," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 14 770–14 780.

[223] Y. Wu *et al.*, "Text2LiDAR: Text-fuided LiDAR point cloud generation via equirectangular transformer," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 291–310.

[224] E. Kirby *et al.*, "LOGen: Toward LiDAR object generation by point diffusion," *arXiv preprint arXiv:2412.07385*, 2024.

[225] T. Yan *et al.*, "OLiDM: Object-aware LiDAR diffusion models for autonomous driving," in *AAAI Conf. Artifi. Intell.*, vol. 39, 2025, pp. 9121–9129.

[226] Y. Xie, C. Xu, C. Peng *et al.*, "X-Drive: Cross-modality consistent multi-sensor data synthesis for driving scenarios," in *Int. Conf. Learn. Represent.*, 2025.

[227] V. Zyrianov *et al.*, "LidarDM: Generative LiDAR simulation in a generated world," *arXiv preprint arXiv:2404.02903*, 2024.

[228] S.-H. Ho, B. Thach, and M. Zhu, "LiDAR-EDIT: LiDAR data generation by editing the object layouts in real-world scenes," in *IEEE Int. Conf. Robot. Autom.*, 2025.

[229] Y. Wu *et al.*, "WeatherGen: A unified diverse weather generator for LiDAR point clouds via spider mamba diffusion," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 17 019–17 028.

[230] T. Martyniuk *et al.*, "LiDPM: Rethinking point diffusion for LiDAR scene completion," in *IEEE Intell. Veh. Symp.*, 2025.

[231] H. Cao and S. Behnke, "DiffSSC: Semantic LiDAR scan completion using denoising diffusion probabilistic models," *arXiv preprint arXiv:2409.18092*, 2024.

[232] X. Zhou, D. Liang, S. Tu *et al.*, "HERMES: A unified self-driving world model for simultaneous 3D scene understanding and generation," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[233] Y. Du *et al.*, "SuperPC: A single diffusion model for point cloud completion, upsampling, denoising, and colorization," *arXiv preprint arXiv:2503.14558*, 2025.

[234] A. Zhao *et al.*, "Diffusion distillation with direct preference optimization for efficient 3D LiDAR scene completion," *arXiv preprint arXiv:2504.11447*, 2025.

[235] C. Shi *et al.*, "DriveX: Omni scene modeling for learning generalizable world knowledge in autonomous driving," *arXiv preprint arXiv:2505.19239*, 2025.

[236] SenseTime-FVG, "Open driving world models (OpenDWM)," https://github.com/SenseTime-FVG/OpenDWM, 2025.

[237] Z. Chen *et al.*, "AnchorFormer: Point cloud completion from discriminative nodes," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 13 581–13 590.

[238] B. Poole *et al.*, "DreamFusion: Text-to-3D using 2D diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[239] B. Wallace, M. Dang, R. Rafailov *et al.*, "Diffusion model alignment using direct preference optimization," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 8228–8238.

[240] H. Chang, H. Zhang, L. Jiang *et al.*, "MaskGIT: Masked generative image transformer," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 11 315–11 325.

[241] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," *arXiv preprint arXiv:2001.10773*, 2020.

[242] X. Wang, Z. Zhu, W. Xu *et al.*, "OpenOccupancy: A large-scale benchmark for surrounding semantic occupancy perception," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17 850–17 859.

[243] Y. Li, S. Li, X. Liu *et al.*, "SSCBench: A large-scale 3D semantic scene completion benchmark for autonomous driving," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024.

[244] Y. Wang *et al.*, "DrivingDojo dataset: Advancing interactive and knowledge-enriched driving world model," *arXiv preprint arXiv:2410.10738*, 2024.

[245] X. Han, Z. Jia, B. Li *et al.*, "Extrapolated urban view synthesis benchmark," *arXiv preprint arXiv:2412.05256*, 2024.

[246] M. Heusel, H. Ramsauer, T. Unterthiner *et al.*, "GANs trained by a two-time-scale update rule converge to a local Nash equilibrium," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[247] T. Unterthiner *et al.*, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[248] D. W. Shu, S. W. Park, and J. Kwon, "3D point cloud generative adversarial network based on tree structured graph convolutions," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3859–3868.

[249] T. Salimans *et al.*, "Improved techniques for training GANs," *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[250] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.

[251] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[252] Z. Huang, Y. He, J. Yu *et al.*, "VBench: Comprehensive benchmark suite for video generative models," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 21 807–21 818.

[253] W. Cao *et al.*, "Pseudo-simulation for autonomous driving," *arXiv preprint arXiv:2506.04218*, 2025.

[254] Y. Hu, J. Yang, L. Chen *et al.*, "Planning-oriented autonomous driving," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 17 853–17 862.

[255] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *IEEE Int. Conf. Robot. Autom.*, 2023, pp. 2774–2781.

[256] S. Wang, Y. Liu, T. Wang *et al.*, "Exploring object-centric temporal modeling for efficient multi-view 3D object detection," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3621–3631.

[257] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 13 760–13 769.

[258] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.

[259] Z. Wang *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[260] R. Zhang, P. Isola, A. A. Efros *et al.*, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 586–595.

[261] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[262] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *Eur. Conf. Comput. Vis. Worksh.*, vol. 90, 2006.

[263] S. Hu, L. Chen, P. Wu *et al.*, "ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 533–549.

[264] W. Tong *et al.*, "Scene as occupancy," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8406–8415.

[265] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[266] Z. Yang *et al.*, "UniSim: A neural closed-loop sensor simulator," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 1389–1399.

[267] Z. Zhang, A. Liniger, D. Dai *et al.*, "TrafficBots: Towards world models for autonomous driving simulation and motion prediction," in *IEEE Int. Conf. Robot. Autom.*, 2023, pp. 1522–1529.

[268] S. Huang, Z. Wang, P. Li *et al.*, "Diffusion-based generation, optimization, and planning in 3D scenes," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 16 750–16 761.

[269] A. Hu, G. Corrado, N. Griffiths *et al.*, "Model-based imitation learning for urban driving," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 20 703–20 716.

[270] M. Zhou, J. Luo, J. Villella *et al.*, "SMARTS: Scalable multi-agent reinforcement learning training school for autonomous driving," in *Conf. Robot Learn.*, 2020.

[271] Y. Jiang *et al.*, "Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities," *arXiv preprint arXiv:2503.05652*, 2025.

[272] A. Szot, A. Clegg, E. Undersander *et al.*, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 251–266.

[273] F. Xia *et al.*, "Gibson env: real-world perception for embodied agents," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 9068–9079.

[274] R. Firoozi, J. Tucker, S. Tian *et al.*, "Foundation models in robotics: Applications, challenges, and the future," *Int. J. Robot. Research*, vol. 44, no. 5, pp. 701–739, 2025.

[275] C. Huang *et al.*, "Visual language maps for robot navigation," *arXiv preprint arXiv:2210.05714*, 2022.

[276] X. Ma, D. Hsu, and W. S. Lee, "Learning latent graph dynamics for visual manipulation of deformable objects," in *IEEE Int. Conf. Robot. Autom.*, 2022, pp. 8266–8273.

[277] D. M. Bear *et al.*, "Physion: Evaluating physical prediction from vision in humans and machines," *arXiv preprint arXiv:2106.08261*, 2021.

[278] K. Mo *et al.*, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 909–918.

[279] R. K. Jones, T. Barton, X. Xu *et al.*, "ShapeAssembly: Learning to generate programs for 3D shape structure synthesis," *ACM Trans. Graphics*, vol. 39, no. 6, pp. 1–20, 2020.

[280] K. Ehsani, W. Han, A. Herrasti *et al.*, "Manipulathor: A framework for visual object manipulation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 4497–4506.

[281] B. Shen, F. Xia, C. Li *et al.*, "iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 7520–7527.

[282] N. Ahmed *et al.*, "A systemic survey of the omniverse platform and its applications in data generation, simulation and metaverse," *Frontiers Computer Sci.*, vol. 6, p. 1423129, 2024.

[283] T. Merino, M. Charity, and J. Togelius, "Interactive latent variable evolution for the generation of minecraft structures," in *Int. Conf. Foundation Digital Games*, 2023, pp. 1–8.

[284] M. Deitke, E. VanderBilt, A. Herrasti *et al.*, "ProcTHOR: Large-scale embodied AI using procedural generation," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 5982–5994.

[285] M. Jia *et al.*, "MGVQ: Could VQ-VAE beat VAE? a generalizable tokenizer with multi-group quantization," *arXiv preprint arXiv:2507.07997*, 2025.

[286] M. Hu *et al.*, "Text2World: Benchmarking large language models for symbolic world model generation," *arXiv preprint arXiv:2502.13092*, 2025.

[287] Z. Ge *et al.*, "WorldGPT: Empowering LLM as multimodal world model," in *ACM Int. Conf. Multimedia*, 2024, pp. 7346–7355.

[288] K. Li *et al.*, "Immersive neural graphics primitives," *arXiv preprint arXiv:2211.13494*, 2022.

[289] Y. Yang, F.-Y. Sun, L. Weihs *et al.*, "HoloDeck: Language guided generation of 3D embodied AI environments," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 16 227–16 237.

[290] C. Gao *et al.*, "Dynamic view synthesis from dynamic monocular video," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5712–5721.

[291] B. Attal, J.-B. Huang, C. Richardt *et al.*, "HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 16 610–16 620.

[292] J. Bai *et al.*, "RecamMaster: Camera-controlled generative rendering from a single video," *arXiv preprint arXiv:2503.11647*, 2025.

[293] J. Yu *et al.*, "GameFactory: Creating new games with generative interactive videos," *arXiv preprint arXiv:2501.08325*, 2025.

[294] S. Gao *et al.*, "AdaWorld: Learning adaptable world models with latent actions," *arXiv preprint arXiv:2503.18938*, 2025.

[295] L. Ying *et al.*, "Assessing adaptive world models in machines with novel games," *arXiv preprint arXiv:2507.12821*, 2025.

[296] L. Lin, Y. Liu, Y. Hu *et al.*, "Capturing, reconstructing, and simulating: the UrbanScene3D dataset," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 93–109.

[297] H. Xie *et al.*, "Generative gaussian splatting for unbounded 3D city generation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 6111–6120.

[298] Y. Shang *et al.*, "UrbanWorld: An urban world model for 3D city generation," *arXiv preprint arXiv:2407.11965*, 2024.

[299] Z. Tang, M. Naphade, M.-Y. Liu *et al.*, "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8797–8806.

[300] S. Tan, J. Lambert, H. Jeon *et al.*, "SceneDiffuser++: City-scale traffic simulation via a generative world model," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 1570–1580.

[301] J. Deng *et al.*, "CityCraft: A real crafter for 3D city generation," *arXiv preprint arXiv:2406.04983*, 2024.

[302] X. Dai, C. Zhao, X. Wang *et al.*, "Image-based traffic signal control via world models," *Frontiers Info. Tech. Electro. Engineer.*, vol. 23, no. 12, pp. 1795–1813, 2022.

[303] J. Hu, X. Dai, X. Li *et al.*, "TrafficWise: Leveraging world models for generalized and interpretable traffic control," *IEEE Intell. Transport. Syst. Magazine*, pp. 2–12, 2025.

[304] F.-Y. Wang, "New control paradigm for industry 5.0: From big models to foundation control and management," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 8, pp. 1643–1646, 2023.

[305] L. Guan *et al.*, "Leveraging pre-trained large language models to construct and utilize world models for model-based task planning," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 79 081–79 094.

[306] J. Pathak *et al.*, "FourcastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators," *arXiv preprint arXiv:2202.11214*, 2022.

[307] K. Bi, L. Xie, H. Zhang *et al.*, "Accurate medium-range global weather forecasting with 3D neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.

[308] R. Lam, A. Sanchez-Gonzalez, M. Willson *et al.*, "Learning skillful medium-range global weather forecasting," *Science*, vol. 382, no. 6677, pp. 1416–1421, 2023.

[309] Y. Yang *et al.*, "Medical world model: Generative simulation of tumor evolution for treatment planning," *arXiv preprint arXiv:2506.02327*, 2025.

[310] B. Kim and J. C. Ye, "Diffusion deformable model for 4D temporal medical image generation," in *Int. Conf. Medical Image Comput. Computer-Assisted Intervention*. Springer, 2022, pp. 539–548.

[311] C. Boton *et al.*, "4D simulation research in construction: A systematic mapping study," *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2451–2472, 2023.

[312] S. Verykokou *et al.*, "3D reconstruction of disaster scenes for urban search and rescue," *Multimedia Tools Appl.*, vol. 77, no. 8, pp. 9691–9717, 2018.

[313] D. Zheng *et al.*, "VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness," *arXiv preprint arXiv:2503.21755*, 2025.

[314] D. Li *et al.*, "Worldmodelbench: Judging video generation models as world models," *arXiv preprint arXiv:2502.20694*, 2025.

[315] H. Duan *et al.*, "WorldScore: A unified evaluation benchmark for world generation," *arXiv preprint arXiv:2504.00983*, 2025.

[316] C. Riquelme, J. Puigcerver, B. Mustafa *et al.*, "Scaling vision with sparse mixture of experts," in *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8583–8595.

[317] H. Chen *et al.*, "Accelerating diffusion models with parallel sampling: Inference at sub-linear time complexity," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 133 661–133 709.

[318] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6299–6308.

[319] A. Milioto *et al.*, "RangeNet++: Fast and accurate LiDAR semantic segmentation," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 4213–4220.

[320] M. Bińkowski *et al.*, "Demystifying MMD GANs," *arXiv preprint arXiv:1801.01401*, 2018.

[321] J. Ke *et al.*, "MUSIQ: Multi-scale image quality transformer," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5148–5157.

[322] J. Sun, Z. Shen, Y. Wang *et al.*, "LoFTR: Detector-free local feature matching with transformers," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 8922–8931.

[323] L. Yang, B. Kang, Z. Huang *et al.*, "Depth anything v2," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 21 875–21 911.

[324] W.-C. Hung *et al.*, "LET-3D-AP: Longitudinal error tolerant 3D average precision for camera-only 3D detection," in *IEEE Int. Conf. Robot. Autom.*, 2024, pp. 8272–8279.

[325] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[326] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 194–210.

[327] X. Weng *et al.*, "3D multi-object tracking: A baseline and new evaluation metrics," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10 359–10 366.

[328] S. Song, F. Yu, A. Zeng *et al.*, "Semantic scene completion from a single depth image," in *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1746–1754.