InfiniVerse: Occupancy Guided Unbounded Scene Generation for Autonomous Driving

Xiaoyu Ye[1,4*,†], Leheng Li[2,4*,†], Xinyu Ji[3*], Yingjie Cai[4], Hongda He[5], Xu Yan[4‡], Guanyi Zhao[4], Ying-Cong Chen[2], Bingbing Liu[4], Shuguang Cui[1], and Zhen Li[1‡]

[1] The Chinese University of Hong Kong, Shenzhen
xiaoyuye1@link.cuhk.edu.cn
{shuguangcui,lizhen}@cuhk.edu.cn
[2] The Hong Kong University of Science and Technology (Guangzhou)
[3] Nanyang Technological University
[4] Huawei Technologies Co., Ltd.
[5] University of New South Wales

Abstract. Generating realistic, controllable, and temporally coherent urban environments is a critical yet unresolved challenge in the autonomous driving community. In this paper, we introduce **InfiniVerse**, a unified pipeline for long-range, 2D–3D-aligned, and controllable synthesis of dynamic urban scenes from a single frame. In practice, our approach first reconstructs a 3D occupancy representation from the input multi-view frame. This representation serves as a foundation for autoregressive scene extension along arbitrary trajectories. Subsequently, a video diffusion model translates the coarse occupancy grid into realistic, spatiotemporally consistent video sequences. Moreover, we propose a hierarchical *sketch-and-refine* paradigm, in which the generated videos are re-projected as image-conditioned feedback to enhance the 3D occupancy representation, establishing cross-modal alignment and mutual enhancement between the visual and spatial domains. Extensive evaluations on the Waymo Open Dataset and nuScenes demonstrate that InfiniVerse achieves state-of-the-art performance, with a FID of **6.4** and FVD of **67.97**, significantly outperforming existing benchmarks in both duration and stability.

Keywords: Scene Generation, Video Generation, Autonomous Driving

1 Introduction

The generation of realistic and controllable synthetic environments is a fundamental challenge in computer vision, with critical applications in autonomous driving, embodied AI, and digital twins [8, 17, 23, 33, 48, 54, 58]. In the context of

*Equal contribution.
†This work was done during an internship at Huawei.
‡Corresponding authors.

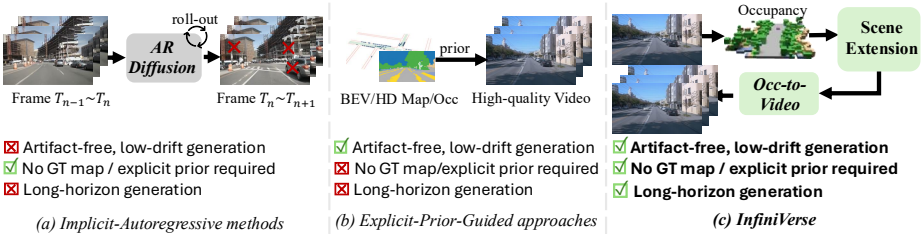


Fig. 1: Paradigms for driving scene generation. Implicit video generation (a) produces visually realistic sequences but lacks explicit 3D structure, often causing geometric drift. Geometry-driven pipelines (b) enforce spatial consistency but rely on strong priors such as BEV/HD maps or occupancy grids. **InfiniVerse (c)** bridges these paradigms by constructing an extensible 3D occupancy world from a single multi-view observation and coupling it with video diffusion through a reciprocal refinement loop, enabling controllable and temporally consistent long-horizon driving video generation.

autonomous driving, the ability to synthesize long-duration, multi-view video sequences is essential for training and validating perception and planning systems. However, generating long-horizon driving scenes remains a significant challenge. Unlike short-range clips, long-term generation requires maintaining strict temporal consistency, physical plausibility, and geometric stability over extended trajectories, where even minor errors can quickly compound into significant visual distortions or structural collapses.

Existing generative approaches for driving scenes can be broadly categorized into two groups, yet both face limitations in long-horizon scenarios. The first category consists of **Implicit-Autoregressive methods** [11,12,44], which typically take a short video clip or a few frames as input and use video diffusion model to autoregressively predict future frames, as depicted in Fig. 1(a). While these models can produce high-fidelity appearance and plausible motion, they often rely on implicit latent features without an explicit intermediate geometric representation. Consequently, during long-term prediction, these methods are prone to accumulated errors and "hallucinations", where the generated content gradually deviates from the underlying 3D structure or loses spatial coherence. The second category focuses on **Explicit-Prior-Guided approaches** [9, 10, 16, 20, 24, 26, 43, 45] (i.e., Fig. 1(b)), requiring a specific BEV map/HD map as input. Although these provide better structural constraints, they often struggle to render long-horizon video due to the limited range of input prior. Furthermore, maintaining alignment between generated geometry and visual appearance remains a difficult open problem.

To address these limitations, we introduce **InfiniVerse** (see Fig. 1(c)), a unified framework for long-range, controllable synthesis of dynamic urban environments from a single multi-view observation. Unlike prior works that rely on implicit transitions or rigid HD maps, our approach reconstructs an explicit 3D occupancy representation from the initial frame. This representation serves as a structural foundation for autoregressively extending the scene along arbitrary

user-defined trajectories. To ensure high-fidelity results, we employ a video diffusion model to translate the extended occupancy grids into consistent multi-view video sequences.

The core of our method is a bidirectional "sketch-and-refine" loop that establishes a closed-loop feedback between the 3D occupancy and 2D video domains. Specifically, we treat the coarse occupancy as a "sketch" to guide video generation, and then re-project the generated high-fidelity frames back into the 3D space as image-conditioned feedback. By making 2D appearance and 3D geometry mutually constraining, our framework significantly reduces drift and suppresses hallucinations unsupported by the underlying scene structure. This reciprocal mechanism enables the creation of long-horizon, temporally consistent, and geometrically aligned driving scenarios.

In summary, our key contributions are:

- We propose **InfiniVerse**, a unified pipeline that generates geometrically accurate and semantically coherent long-range 3D scenes from a single frame.
- We leverage a hierarchical **sketch-and-refine paradigm** that progressively refines the occupancy representation, enhancing scene consistency and controllability over long trajectories.
- Extensive experiments on the Waymo Open Dataset and nuScenes demonstrate that InfiniVerse achieves new state-of-the-art performance in both generation quality and duration.

2 Related work

2.1 3D Urban Scene Generation

Recent progress [53] has been driven by learning-based methods that leverage powerful priors from large-scale 3D datasets, enabling diffusion model based 3D reconstructions [29, 30] even under severe ambiguity. Due to lack of high-quality 3D assets, several simulated environments played a vital role in autonomous driving, with platforms like CARLA [7] and Waymo's Simulation Engine [37] providing handcrafted 3D assets and rule-based control. However, these systems require significant manual effort and lack flexibility for arbitrary trajectory exploration.

More recent methods attempt to automate urban scene generation by leveraging learned priors from real-world datasets. For instance, Infinicube [25] has proposed techniques for large-scale scene synthesis, but they often depend on highly annotated HD maps and 3D layout bounding boxes, which limits scalability and data efficiency. X-Scene [51] explores large-scale driving scene generation guided by textual descriptions, enabling diverse scene composition and condition control through natural language prompts. However, its control mechanism remains coarse-grained and ambiguous, making it difficult to achieve precise spatial or semantic alignment between text instructions and generated content. Some methods [19, 23, 46] have exploited block-by-block generation for semantic scene generation in outdoor environments but lack controllability and fidelity.

2.2 3D Occupancy Representation

CityDreamer [47] learns NeRF-style radiance fields or signed-distance functions to render novel views, but store geometry only implicitly; they cannot expose occupancy grids required for downstream planning or physical simulation.

Occupancy grids have emerged as a powerful and compact representation for 3D scene understanding, especially in autonomous driving and robotics due to its explicit representation for efficient computation. They provide explicit volumetric reasoning and enable consistent spatial modeling across time and views. Works such as OccNet [34], VoxelNeXt [4], and Occ3D [39] have shown promise in using learned occupancy fields for tasks like object detection and motion prediction. However, these approaches are typically limited to static or short-term settings and do not support long-range generation.

2.3 Controllable Video Generation

Video generation has recently seen significant progress with the introduction of scalable and general-purpose generative models. SVD (Stable Video Diffusion) [2] introduces a diffusion-based framework that models video as a sequence of latent frames, enabling high-quality and temporally coherent video generation. Its scalable architecture and strong pretraining enable it to generalize well across diverse prompts and temporal lengths. Building on this, CogVideoX [52] proposes a diffusion transformer video generation framework that compresses videos across spatial and temporal dimensions in a 3D Variational Autoencoder (VAE). By using hierarchical spatiotemporal transformers and a factorized design, Large-scale diffusion models like VideoComposer [42] and CogVideoX [52] condition on text, masks, or optical flow to produce plausible short clips, yet lack any notion of depth, occluded content, or kilometre-scale consistency but provide a very powerful base model.

3 Method

3.1 Overview

Our framework adopts a two-stage design for large-scale 3D scene and long-term video generation. In the first stage, we reconstruct a detailed occupancy representation from a multi-view frame and encode it into a triplane-based generator, enabling fast and flexible synthesis of coarse 3D occupancy scene under arbitrary configurations. In the second stage, we fine-tune a video diffusion model to translate the synthesized occupancy scene into high-fidelity multi-view driving videos. The generated videos are further utilized to refine the occupancy representation, forming a fine-grained updating loop. Throughout the process, we construct reference images and multiple control conditions to ensure long-term temporal and cross-modal consistency between the occupancy representation and video generation.

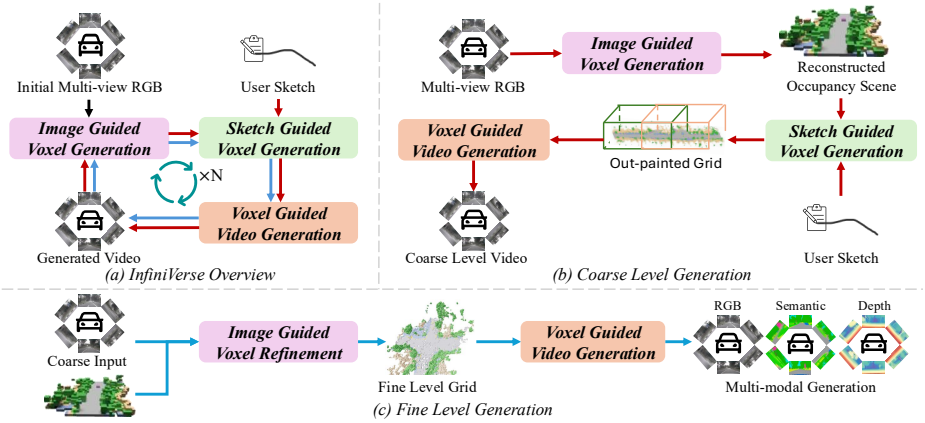


Fig. 2: InfiniVerse overview. We establish the system from a single multi-view frame by: 1) Constructing an image-guided voxel diffusion network to precisely reconstruct current initial occupancy scene and then encode the scene into triplane, performing sketch-guided occupancy generation to create a long-range, coarse level large-scale occupancy scene; and then 2) utilizing a fine-tuned video generator to translate the semantic world into coarse level driving video; 3) re-projecting the generated RGB video into image-conditioned voxel diffusion network to further refine each scene chunk, forming a reciprocal loop to provide 2D-3D aligned large scale scene generation. This scene can then be extended auto-regressively for long horizon exploration.

3.2 Image Guided Voxel Generation

We adopt XCube [29] as the backbone for our initial scene reconstruction module. XCube is a large-scale 3D generative model that represents geometry using hierarchical sparse voxel grids and synthesizes scenes via a latent diffusion process. Specifically, it learns a distribution over a latent variable $\mathbf{X}$ encoded by a sparse structure Variational Autoencoder (VAE) defined on voxelized 3D geometry. The VAE encodes sparse voxel hierarchies into compact latent representations, and a diffusion model operates in this latent space to generate sparse voxel structures. Both the VAE encoder–decoder and the diffusion model are implemented using sparse convolutional neural networks [13], enabling efficient modeling of high-resolution geometry up to 1024^3.

To condition XCube on input images, we use LSS [28] unprojection pipeline to encode the image feature extracted from DINO-V2 [27] into a dense 3D voxel grid $\mathbf{\Omega}$. The image features are processed with trainable 2D conv layers stored into feature channel $\mathbf{C}$

$$\mathbf{F}^i_{jd} = \theta^i_{jd} \cdot \mathbf{F}^i_j, \quad \mathbf{C}_v = \sum_{(i,j,d)} \mathbf{F}^i_{jd} \in \mathcal{R}^C. \tag{1}$$

The $\theta^i_{jd} \in \mathbb{R}^D$ is D-dimensional Softmax-normalized vector pixel level depth distribution, the $d \in [1, D]$ denotes index of the depth pixel where v denotes the index of a voxel. A key difference from indoor or object-centric setups is

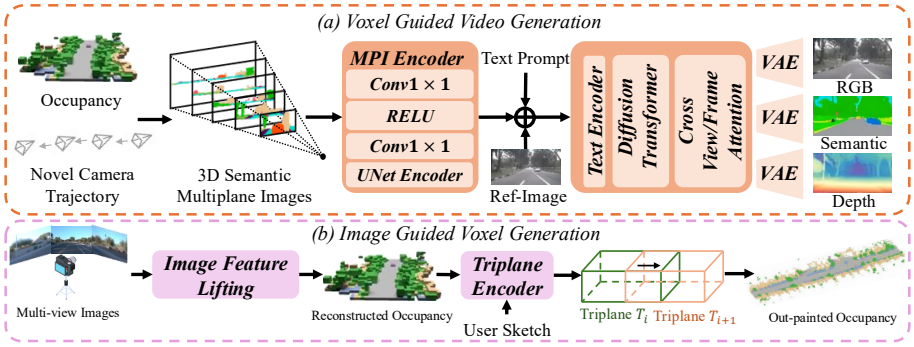


Fig. 3: Detail illustration on Voxel-to-Video diffusion and Video-to-Voxel diffusion. We Construct (a) by first encode the occupancy scene into Multiplane images to efficiently store both the semantic and geometric information, we designed a 1×1 convolutional encoder without downsampling to encode MPI features and transformed by a 1×1 convolution layer and ReLU activation. We also concatenate ref-image and Text conditions to further maintain the consistency and provide further control. With different VAE decoder we managed to output Multi-view RGB, semantics, depth video. As illustration (b), with a snapshot multi-view image, we first encode the image features in occupancy grid and with a Voxel diffusion, we reconstruct detailed occupancy scene and encode it in triplane with sketch condition C_{sketch} for arbitrary sketch guided scene extrapolate.

that outdoor driving scenes provide sparse and uneven camera coverage, we cannot naively broadcast per-pixel features along the entire ray, as done in prior work [22, 35, 36] since that the geometries is not precisely corresponding with camera frusta. Once we constructed condition grid $\mathbf{C}$, we concatenate it with the latent $\mathbf{X}$ and feed into diffusion network as conditioning to produce high quality voxel grid for further extrapolator.

3.3 Sketch Guided Voxel Generation

Our framework represents the environment as an explicit occupancy grid $\mathcal{O}_t \in \{0,1\}^{H\times W\times D}$ that evolves over timesteps t. Given the single-frame multi-view input, the reconstruction module (Sec. 3.2) reconstructs $\mathcal{O}_0$ with fine-grained semantics and geometry. We regard this grid as the initial state of and learn a triplane based diffusion head $q(\mathcal{O}_{t+1} \mid \mathcal{O}_t, \mathbf{u}_t)$ that predicts the next occupancy volume conditioned on (i) the current grid $\mathcal{O}_t$ and (ii) a compact control vector $\mathbf{u}_t$ encoding the desired chunk Sketch trajectory. To address the challenge of maintaining geometric and semantic coherence across extended trajectories, we propose a hierarchical sketch-and-refine paradigm that decomposes long-range occupancy generation into two stages: coarse-level sketching and fine-level image-conditioned refinement method.

To achieve controllable occupancy scene out-painting and fine-level image-conditioned occupancy updating, we encode the semantic occupancy grid to triplane latent representation [19]. The triplane latent projects the 3D occupancy

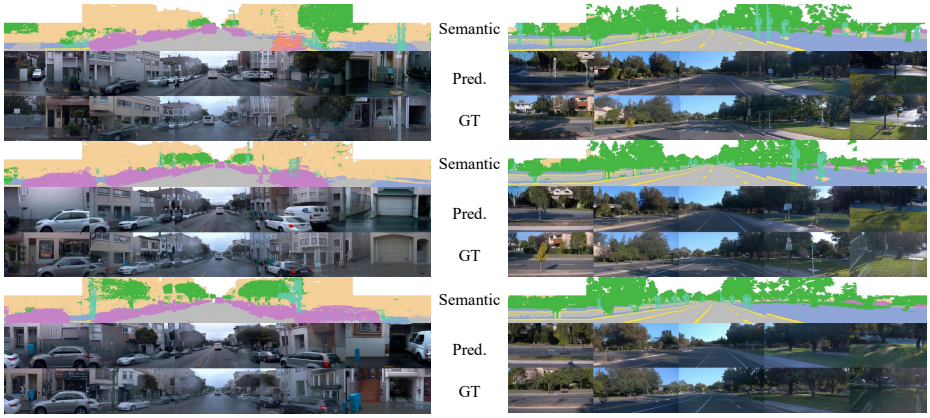


Fig. 4: Top to bottom: generated frames at T+5, T+15, and T+25 on Waymo Open Dataset. The Semantic row denotes Generated occupancy scenes encoded via the MPI encoder, while the ground-truth (GT) frames are shown in the bottom row. The comparison highlights the high fidelity and strong Temporal consistency of our generated sequences across the entire temporal span.

onto three orthogonal rigid axes, decomposing the volumetric representation into three complementary 2D feature planes: XY, XZ, and YZ planes. By querying features from all three planes and aggregating them through learned fusion mechanisms, we can regenerate detailed 3D semantic information while maintaining computational efficiency. In the subsequent parts of this section, all operations are performed within the triplane representation space.

Coarse-level Sketching Guided Generation. Given an initial scene $\mathbf{O}_0$ and a trajectory sequence $\{T_j\}_{j=0}^{N}$, the sketch stage generates a coarse-level scene layout along the entire control trajectory. This stage operates at reduced spatial resolution to capture the global scene structure efficiently. The Sketch generation is conditioned on both the trajectory and the initial scene context for coherence:

$$\mathbf{X}_{sketch} = D_\theta(\mathbf{x}_0, \{T_j\}_{j=0}^{N}), \tag{2}$$

Where D_θ denotes denoising loop, $\mathbf{X}_{sketch}$ represents the generated coarse-level scene spanning the full trajectory with a volume of 128^3. Then we sample the occupancy alone the trajectory with a standard resolution triplane and also a sliding window denoising loop for a coarse level occupancy representation as shown here:

$$\mathbf{X}_i = D_\theta(\mathbf{V}_{i-1}; \mathbf{C}_{context}, \mathbf{C}_{sketch}), \tag{3}$$

where $\mathbf{C}_{context}$ and $\mathbf{C}_{sketch}$ contains details from the previously refined patch $\mathbf{X}_{i-1}$, $\mathbf{V}_i$ indicates the triplane spatial rotation and translation vector for triplane to sample alone the trajectory. This step ensures smooth transitions and maintains local coherence across window boundaries.

This step provides a semantically consistent global layout that serves as the foundation for subsequent refinement.

3.4 Bidirectional Closed-Loop Generation

Controllable Video Generation. We use CogvideoX1.5-5B as our base model and LoRA-fine-tuned on our processed dataset. It is non-trivial to design a 3D representation for conditioning and generating, but we aim to efficiently store both the semantic and geometric information of any occupancy input. We use multi-plane images (MPIs) [57] as a representation to further improve video quality through acquiring 3D information of occluded objects.

ControlNet-style Injection. As demonstrated in Fig. 3, we encode the D-plane stack acquired from camera frustum with a 1×1 Conv + ReLU encoder ϕ_M (no down-sampling) and inject it into the UNet decoder features $\{h^\ell\}$ at matching resolutions:

$$\tilde{h}^\ell \;\leftarrow\; h^\ell \;+\; \lambda_M \, \phi_M\big(\mathrm{MPI}^\ell\big), \tag{4}$$

where λ_M scales geometry control. Using 1×1 keeps exact spatial correspondence between multi-plane features and latent pixels, improving label alignment and compute over 3×3 alternatives.

Multi-view/frame Consistency. We use zero-initialized cross-view attention to exchange information with neighboring cameras and cross-frame attention to propagate temporal context:

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\Big(\tfrac{QK^\top}{\sqrt{d}}\Big)V, \tag{5}$$

$$h_{\mathrm{out}} = h_{\mathrm{in}} + \sum_{i\in\{l,r\}} \mathrm{Attn}(Q_{\mathrm{in}}, K_i, V_i), \tag{6}$$

$$h_{\mathrm{out}} = h_{\mathrm{in}} + \sum_{i\in\{f,h\}} \mathrm{Attn}(Q_{\mathrm{in}}, K_i, V_i), \tag{7}$$

where $\{l, r\}$ denote left/right adjacent views and $\{f, h\}$ future/history frames. Eqs. (6)–(7) maintain instance coherence across overlapping FoVs and time. We also design a Ref-image initial conditioned injection mechanism for long-term video stability and maintain the temporal consistency. For every video generation chunk, we took last generated video frame or the initial frame if its the first chunk F_t, and encoded it through vanilla VAE proposed from CogvideoX. Then we concat it through similar conditioned injection introduced in Sec 3.4 to formulated a text-pfirompted and ref-image enabled spatial temporal consistent video generation head. As illustrated in Fig. 3 we route different vae decoders for various output. Through this manner, we support 2D-3D aligned video generation with precisely occupancy-centric conditioning and outputs across semantics, depth and RGB demonstrated in Fig. 5 in nuScenes. We also present down-sampled lidar from our generated occupancy grid as lidar sensor output.

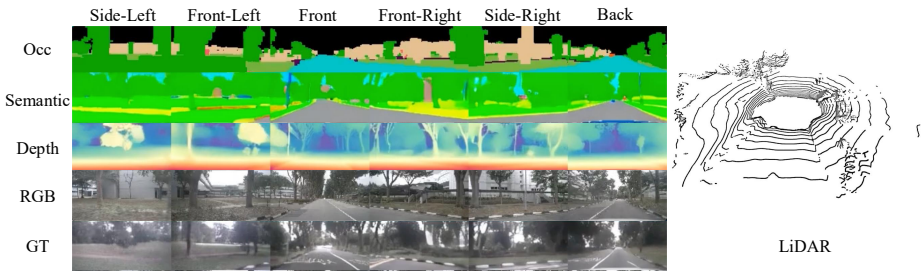


Fig. 5: Demonstration on aligned multiple multi-view sensor output compared with GT RGB Frame on nuScenes.

Fine-level Refinement. The refinement stage operates on local scene patches using a sliding-window strategy to progressively enhance the coarse, sketch-based scene with fine-grained details contributed by the powerful video diffusion head. For each window position i-th, we encode the corresponding camera frustum into a multi-plane image (MPI), which serves as the 3D condition for a fine-tuned video diffusion model. This process generates spatially and temporally consistent multi-view videos with high visual fidelity. With these high-quality videos, the large-scale driving scene generation task is now an image-conditioned occupancy reconstruction problem. We thus reuse the reconstruction model introduced in Sec. 3.2 to obtain detailed, high-resolution occupancy representations. This improved occupancy further enables more accurate video generation, forming a closed-loop refinement process that progressively enhances both the 3D occupancy and the visual quality of the driving videos, achieving strong multi-modal alignment.

This two-stage approach enables the generation of arbitrarily long scenes while preserving both global semantic consistency through the sketch stage and local geometric fidelity through the refinement stage. The overlapping regions between adjacent windows provide continuity constraints that prevent artifacts and maintain seamless scene transitions throughout the extended trajectory.

Text-driven Weather and Style Control. High-level appearance is provided by a text encoder $\tau_\theta(\cdot)$ that embeds prompts y into conditioning tokens $g = \tau_\theta(y)$. We form y by concatenating weather/style clauses (e.g., "snowy", "sandstorm") with scene text; classifier-free guidance is used at sampling time. We inject g via the standard cross-attention layers of the latent diffusion UNet.

Unified Denoising Objective. We train the video diffusion model using a standard latent denoising objective conditioned on all available control signals:

$$\mathcal{L}_{\text{base}} \;=\; \mathbb{E}_{\epsilon,\,t}\Big[\big\|\epsilon - \epsilon_\theta(z_t,\,t,\,\phi_M(\text{MPI}))\,\big\|_2^2 \odot w\Big], \tag{8}$$

where ϵ_θ denotes the noise predicted by the denoising network at diffusion step t, z_t is the noisy latent, and $\phi_M(\text{MPI})$ encodes the multi-plane image (MPI) as the

3D control signal. The weight term w optionally applies importance weighting, such as progressive or depth-aware weighting, to stabilize training and emphasize geometrically meaningful regions.

To further enhance the generation quality of dynamic scene components, we introduce an additional region-weighted reconstruction loss that assigns higher importance to road surfaces and road actors (vehicles, pedestrians, cyclists, et al.):

$$\mathcal{L}_{\text{region}} \;=\; \mathbb{E}_{\epsilon,\,t}\Big[\big\|\big(\epsilon - \epsilon_\theta(z_t,\,t,\,\phi_M(\text{MPI}))\,\big)\odot m_{\text{road}}\big\|_2^2\Big], \tag{9}$$

where m_{road} is a binary or soft mask highlighting road and actor regions derived from semantic category.

The final training objective combines both terms:

$$\mathcal{L}_{\text{video}} \;=\; \mathcal{L}_{\text{base}} \;+\; \lambda_{\text{region}}\,\mathcal{L}_{\text{region}}, \tag{10}$$

where λ_{region} controls the relative importance of the region-weighted constraint.

This composite loss encourages the diffusion model to focus on both global temporal coherence and fine-grained generation of critical driving scene elements, yielding visually consistent and semantically accurate multi-view videos and also improves the fidelity of the assigned category in the bidirectional cross-Modal closed-loop generation.

Ultra-long video rollout. Let $\{x_{1:F}\}$ be the target frame sequence, and let $\mathcal{C} = \{\text{MPI}, g\}$ be the fixed controls for the rollout. We generate with a sliding-window sampler of length W using cross-frame attention:

$$p_\theta(x_t \mid x_{t-W:t-1}, \mathcal{C}) \;\propto\; \prod_{k=K}^{1} p_\theta\big(z_{k-1} \mid z_k,\,\mathcal{C},\,x_{t-W:t-1}\big), \tag{11}$$

periodically re-anchoring features with $\phi_M(\text{MPI})$ (every R frames) to curb drift and preserve long-horizon geometric fidelity. The cross-frame operator in equation (7) enables temporal propagation beyond training clip lengths; combined with prompt controls, it sustains consistent appearance over arbitrarily long sequences.

4 Experiment

4.1 Dataset and Experimental Setup

Data Curation We conduct comprehensive experiments on the nuScenes [3] and Waymo Open Dataset [37]. To support training and evaluation, we further extend the data processing pipeline provided by SCube [30] to construct a more accurate 3D occupancy dataset. To enhance the geometric fidelity of our representations, we integrate dense geometry reconstructed by the multi-view stereo pipeline of COLMAP [32], which significantly enriches the voxel representation

with fine-grained structural details and effectively addresses the height limitations inherent in LiDAR sensors. We further add dynamic road actors extracted from ground-truth LiDAR frames [39], including vehicles, pedestrians, and cyclists et al. A key focus of this data curation process is maintaining precise alignment across modalities—ensuring that RGB imagery, LiDAR point clouds, and occupancy representations are geometrically and temporally consistent within a unified world coordinate frame. This alignment is crucial for learning reliable 2D–3D correspondences, enabling InfiniVerse to jointly reason about appearance, geometry, and dynamics. Each curated scene therefore contains both static environmental geometry and temporally coherent dynamic elements, providing a strong foundation for long-horizon, controllable, and multi-modal grounded scene generation. To enable precise text control for video generation, we implement a sophisticated multi-stage captioning process. First, we divide each video clip into three segments corresponding to different front-view perspectives. For each segment, we apply a multi-stage captioning process across three front-views, beginning with InternVL2-40B-AWQ [5, 6] to generate dense video captions. These captions are then refined using Meta-Llama-3-8B-Instruct [1] to improve linguistic quality and ensure better alignment with the requirements of controllable video generation. To guarantee semantic consistency between video and text, we further apply VideoCLIP-XL [41] to filter out poorly aligned video-caption pairs.

Following standard practices in the field, we remove low-quality voxel grids and sequences with predominantly static ego trajectories, resulting in a curated dataset of 618 sequences for training and 130 sequences for evaluation in Waymo Open Dataset. We also use Occ3D-nuScenes [39] to train the occupancy generation branch and nuScenes [3] to train the multi-view multi-modal video generation branch. This filtering and dataset selection ensures the quality and relevance of our experimental data.

4.2 Implementation Details

Our implementation follows a carefully designed training protocol to ensure optimal performance across all components. The image-conditioned occupancy generation stage is trained for 100 GPU days including VAE compression, with a inference time in 2 minutes and 40GB VRAM per chunk. We use four 2D

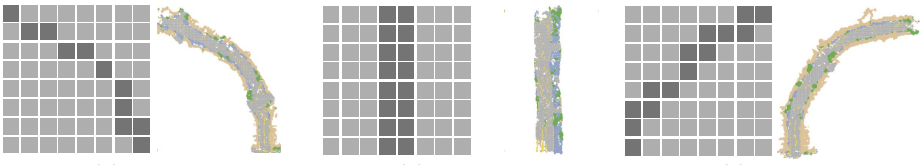


Fig. 6: Conditions and results for sketch guided voxel generation, we provide generated left turn, straight lane and right turn scenario based on corresponding sketch.

Table 1: Comparison of input/output between previous method and ours. Our method produces aligned long video starts by only one frame of input. With minimal input we achieve a new state-of-the-art result in FID(↓) in nuScenes validation set. "MV video" denotes Multi-view video generation.

Method	Input Type	Output Type		2D-3D Aligned	FID↓
		MV Video	3D Repre.		
BEVGen [38]	BEV Map	×	×	-	40.48
BEVControl [50]	BEV Sketch	×	×	-	29.04
GenAD [49]	Image+BEV map	×	×	-	23.90
Drive-WM [44]	Image+Action+Box/Map	✓	×	-	15.8
DriveDreamer-2 [55]	Box+FoV Map	✓	HD Map+BEV	×	25.0
MagicDrive [11]	Pose+Box+BEV	✓	×	-	16.59
MagicDrive-V2 [10]	Traj.+Box+BEV	✓	×	-	20.91
MagicDrive-3D [9]	Traj.+Box+BEV	✓	3DGS	×	20.67
InfiniCity [21]	Voxel	×	Voxel	×	77.0
Panacea [45]	Image+Box+FoV Map	✓	×	-	16.87
Vista [12]	Action	×	×	-	13.97
Uniscene [20]	BEV	✓	Voxel+LiDAR	×	6.45
DiVE [16]	Traj.+Box+FoV	✓	×	-	10.68
DelPhi [26]	BEV Map	✓	×	-	15.08
X-Scene [51]	Text+Box+HD Map	✓	Voxel	×	12.77
InfiniCube [25]	Box+HD Map	×	Voxel+3DGS	×	80.13
DriveGan [18]	Image+Action	×	×	-	73.4
WovoGen [24]	Image+HD Map+Voxel	✓	HD Map+Voxel	×	27.6
DriveDreamer [43]	Image+Action+FoV Map	✓	×	-	14.9
InfiniVerse (ours)	**Image**	✓	Voxel+LiDAR	✓	**6.40**

convolutional layers(channel dims: [768, 256, 256, 32, 32], kernel size: 3, stride: 1) to further process the DINO-V2 output to predict the image feature. The triplane based sketch guided diffusion is trained for 20 GPU days with inference time in 30s per triplane. For the triplane autoencoder, the input scene is encoded to triplane with a spatial resolution $(X_h, Y_h, Z_h) = (128, 128, 32)$, and the feature dimension C_h is 16, the learning rate is initialized to 1e-4 and then decreases linearly. During the diffusion process, we use the default settings [31] with 100 time steps (T). The video generation component is trained 64 GPU days, with a inference time in 2.5 minutes with 55GB VRAM per chunk. We adopt CogvideoX1.5-5B as our base model and LoRA-fine-tuned on our processed dataset. We use UniPC scheduler [56] with the classifier-free guidance (CFG) [15] that is set as 7.0. During inference, we use 20 denoising steps for dataset generation. We employ a temporal compression ratio of 4 to balance computational efficiency with temporal fidelity. All experiments are conducted on NVIDIA A800 GPUs with 80GB VRAM with mixed precision training to optimize memory usage and training speed.

Table 2: Ablation of architectural components for video generation. We analyze the contribution of the 3D branch, the 2D–3D reciprocal loop, and the region-aware loss under the presence of a reference image. Without 3D branch the experiment is conducted under pure I2V mode. FID(↓) and FVD(↓) demonstrate that each component significantly improve generation quality.

Ref-Image	3D branch	Reciprocal Loop	Region Loss	FID↓	FVD↓
×	✓	×	×	40.87	179.4
✓	×	×	×	47.89	320.6
✓	✓	×	×	19.12	142.12
✓	✓	✓	×	7.98	80.87
✓	✓	✓	✓	**6.40**	**67.97**

Table 3: FVD under different rollout lengths. Lower is better.

Method \ Frames	12	24	36	48	96
Vista	78.4	86.2	94.7	105.9	141.3
Ours	**52.1**	**59.8**	**67.5**	**73.9**	**86.5**

4.3 Metrics

We use Fréchet Inception Distance (FID) [14] and Fréchet Video Distance (FVD) [40] to measure the perceptual quality of generated video, and use mIoU to measure the precision of occupancy prediction.

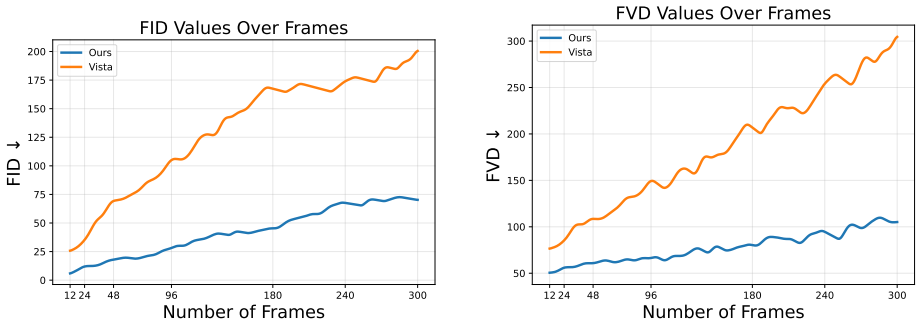
For Voxel grid reconstruction, we utilize IoU and mIoU, the final IoU of the fine-level voxel grid reached 34.31%, and the mIoU that considers the accuracy of the voxel's semantic prediction reached 20.12%.

4.4 Results

Fig. 4 presents qualitative results of our out-painted driving world together with the corresponding RGB video generation on the Waymo Open Dataset [37]. From top to bottom, we visualize the generated future frames along the predicted trajectory. The first frame corresponds to the ground-truth observation, which provides the initial appearance conditions such as weather and scene style. The subsequent frames illustrate that our method can generate long-horizon driving videos with realistic appearance while maintaining strong controllability.

Fig. 5 further demonstrates the multi-modal consistency of our framework on the nuScenes dataset [3]. Our generated results are aligned across multiple modalities, including RGB images, semantics, depth, occupancy, and LiDAR, highlighting the effectiveness of the proposed 2D–3D coupled generation paradigm.

Fig. 6 shows that, given an arbitrary user-defined road layout sketch, our framework generates a long-range occupancy scene that follows the specified layout and guides the subsequent video generation. Tab. 1 provides a quantitative



(a) FID values over frames on Waymo. **(b)** FVD values over frames on nuScenes.

Fig. 7: Long-horizon generation quality under increasing rollout horizons. InfiniVerse shows a substantially slower degradation trend than Vista in both frame-level fidelity and temporal consistency.

comparison with existing methods. Despite using only a single multi-view frame as input, InfiniVerse achieves state-of-the-art video generation quality. Thanks to the proposed reciprocal refinement loop, our framework is able to produce the first long-horizon driving videos that remain consistently aligned with the underlying 3D occupancy representation.

We report a horizon-dependent FID analysis on Waymo. Since Vista [12] does not support multi-view generation, we perform this comparison in the single-view setting. For each target horizon T, we evaluate the generated frame at step $T+1$ and plot FID as a function of rollout length up to 300 frames.

We compare InfiniVerse against Vista in this analysis. Vista is re-implemented by ourselves. As shown in Fig. 7a, the FID of InfiniVerse increases much more slowly than that of Vista as the rollout horizon becomes longer. As shown in Fig. 7b, we report FVD at rollout horizons of 12, 24, 36, 48, 96, and up to 300 frames. As the rollout horizon increases, InfiniVerse consistently maintains lower FVD values and exhibits a slower degradation trend, indicating stronger long-range temporal stability. Table 3 presents a direct FVD comparison with Vista [12] under different rollout lengths. Among existing baselines, Vista is the only method that can be both reproduced and adapted to a setting close to ours.

We further conduct extensive ablation studies to evaluate the effectiveness of each component in the proposed pipeline in Tab. 2. Using only the 3D branch (voxel-to-video) or only the image branch (image-to-video) results in noticeable performance degradation due to the lack of either temporal consistency or geometric priors. Starting from the reference image and incorporating the 3D occupancy representation, the introduction of the proposed reciprocal loop and region loss significantly improves generation quality, validating the effectiveness of our design.

5 Conclusion

In this paper, we introduced **InfiniVerse**, a unified framework for controllable long-horizon generation of dynamic urban scenes for autonomous driving. Starting from a single multi-view observation, our method reconstructs a 3D occupancy representation that serves as a geometric foundation for autoregressive scene expansion along arbitrary trajectories.

To bridge spatial reasoning and visual realism, we couple occupancy-guided world modeling with a video diffusion generator. Central to our framework is a *sketch-and-refine* paradigm that establishes a reciprocal feedback loop between 3D occupancy generation and 2D video synthesis: coarse occupancy grids provide structural guidance for video generation, while the generated frames are reprojected to refine the underlying 3D scene representation.

This bidirectional interaction enables consistent geometry–appearance alignment and significantly improves long-horizon stability while reducing hallucination artifacts. Extensive experiments demonstrate that InfiniVerse generates temporally consistent and visually realistic driving scenarios, achieving state-of-the-art performance on standard benchmarks.

Overall, InfiniVerse provides a scalable paradigm for synthesizing controllable urban environments, offering a promising direction for data generation, simulation, and large-scale evaluation in autonomous driving systems.

References

1. AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023), https://arxiv.org/abs/2311.15127
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving (2020), https://arxiv.org/abs/1903.11027
4. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking (2023), https://arxiv.org/abs/2303.11301
5. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
7. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator (2017), https://arxiv.org/abs/1711.03938
8. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models (2024), https://arxiv.org/abs/2405.10314
9. Gao, R., Chen, K., Li, Z., Hong, L., Li, Z., Xu, Q.: Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes (2025), https://arxiv.org/abs/2405.14475
10. Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control (2025), https://arxiv.org/abs/2411.13807
11. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: MagicDrive: Street view generation with diverse 3d geometry control. In: International Conference on Learning Representations (2024)
12. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability (2024), https://arxiv.org/abs/2405.17398
13. Graham, B., van der Maaten, L.: Submanifold sparse convolutional networks (2017), https://arxiv.org/abs/1706.01307
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018), https://arxiv.org/abs/1706.08500
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance (2022), https://arxiv.org/abs/2207.12598
16. Jiang, J., Hong, G., Zhang, M., Hu, H., Zhan, K., Shao, R., Nie, L.: Dive: Efficient multi-view driving scenes generation based on video diffusion transformer (2025), https://arxiv.org/abs/2504.19614
17. Kim, S.W., Brown, B., Yin, K., Kreis, K., Schwarz, K., Li, D., Rombach, R., Torralba, A., Fidler, S.: Neuralfield-ldm: Scene generation with hierarchical latent diffusion models (2023), https://arxiv.org/abs/2304.09787

18. Kim, S.W., Philion, J., Torralba, A., Fidler, S.: Drivegan: Towards a controllable high-quality neural simulation (2021), https://arxiv.org/abs/2104.15060
19. Lee, J., Lee, S., Jo, C., Im, W., Seon, J., Yoon, S.E.: Semcity: Semantic scene generation with triplane diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2024)
20. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. arXiv preprint arXiv:2412.05435 (2024)
21. Lin, C.H., Lee, H.Y., Menapace, W., Chai, M., Siarohin, A., Yang, M.H., Tulyakov, S.: Infinicity: Infinite-scale city synthesis (2023), https://arxiv.org/abs/2301.09637
22. Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion (2023), https://arxiv.org/abs/2311.07885
23. Liu, Y., Li, X., Li, X., Qi, L., Li, C., Yang, M.H.: Pyramid diffusion for fine 3d large scene generation (2024), https://arxiv.org/abs/2311.12085
24. Lu, J., Huang, Z., Yang, Z., Zhang, J., Zhang, L.: Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation (2024), https://arxiv.org/abs/2312.02934
25. Lu, Y., Ren, X., Yang, J., Shen, T., Wu, Z., Gao, J., Wang, Y., Chen, S., Chen, M., Fidler, S., Huang, J.: Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models (2024), https://arxiv.org/abs/2412.03934
26. Ma, E., Zhou, L., Tang, T., Zhang, Z., Han, D., Jiang, J., Zhan, K., Jia, P., Lang, X., Sun, H., Lin, D., Yu, K.: Unleashing generalization of end-to-end autonomous driving with controllable long video generation (2024), https://arxiv.org/abs/2406.01349
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
28. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d (2020), https://arxiv.org/abs/2008.05711
29. Ren, X., Huang, J., Zeng, X., Museth, K., Fidler, S., Williams, F.: Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
30. Ren, X., Lu, Y., Liang, H., Wu, J.Z., Ling, H., Chen, M., Fidler, Sanja annd Williams, F., Huang, J.: Scube: Instant large-scale scene reconstruction using voxsplats. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), https://arxiv.org/abs/2112.10752
32. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
33. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation (2024), https://arxiv.org/abs/2308.16512
34. Sima, C., Tong, W., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., Li, H.: Scene as occupancy (2023)

35. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhöfer, M.: Deepvoxels: Learning persistent 3d feature embeddings (2019), https://arxiv.org/abs/1812.01024
36. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: Neuralrecon: Real-time coherent 3d reconstruction from monocular video (2021), https://arxiv.org/abs/2104.00681
37. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
38. Swerdlow, A., Xu, R., Zhou, B.: Street-view image generation from a bird's-eye view layout (2024), https://arxiv.org/abs/2301.04634
39. Tian, X., Jiang, T., Yun, L., Mao, Y., Yang, H., Wang, Y., Wang, Y., Zhao, H.: Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving (2023), https://arxiv.org/abs/2304.14365
40. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: FVD: A new metric for video generation. In: Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019. OpenReview.net (2019), https://openreview.net/forum?id=rylgEULtdN
41. Wang, J., Wang, C., Huang, K., Huang, J., Jin, L.: Videoclip-xl: Advancing long description understanding for video clip models (2024), https://arxiv.org/abs/2410.00741
42. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability (2023), https://arxiv.org/abs/2306.02018
43. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-driven world models for autonomous driving (2023), https://arxiv.org/abs/2309.09777
44. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving (2023), https://arxiv.org/abs/2311.17918
45. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving (2023), https://arxiv.org/abs/2311.16813
46. Wu, Z., Li, Y., Yan, H., Shang, T., Sun, W., Wang, S., Cui, R., Liu, W., Sato, H., Li, H., Ji, P.: Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. ACM Transactions on Graphics **43**(4) (2024). https://doi.org/10.1145/3658188
47. Xie, H., Chen, Z., Hong, F., Liu, Z.: Citydreamer: Compositional generative model of unbounded 3d cities (2024), https://arxiv.org/abs/2309.00610
48. Xu, Y., Shi, Z., Yifan, W., Chen, H., Yang, C., Peng, S., Shen, Y., Wetzstein, G.: Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation (2024), https://arxiv.org/abs/2403.14621
49. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., Zhang, J., Geiger, A., Qiao, Y., Li, H.: Genad: Generalized predictive model for autonomous driving (2024), https://arxiv.org/abs/2403.09630
50. Yang, K., Ma, E., Peng, J., Guo, Q., Lin, D., Yu, K.: Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout (2023), https://arxiv.org/abs/2308.01661

51. Yang, Y., Liang, A., Mei, J., Ma, Y., Liu, Y., Lee, G.H.: X-scene: Large-scale driving scene generation with high fidelity and flexible controllability (2025), https://arxiv.org/abs/2506.13558
52. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer (2025), https://arxiv.org/abs/2408.06072
53. Ye, C., Wu, Y., Lu, Z., Chang, J., Guo, X., Zhou, J., Zhao, H., Han, X.: Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. arXiv preprint arXiv:2503.22236 (2025)
54. Zhang, J., Zhang, Q., Zhang, L., Kompella, R.R., Liu, G., Zhou, B.: Urban scene diffusion through semantic occupancy map (2024), https://arxiv.org/abs/2403.11697
55. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation (2024), https://arxiv.org/abs/2403.06845
56. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models (2023), https://arxiv.org/abs/2302.04867
57. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images (2018), https://arxiv.org/abs/1805.09817
58. Zyrianov, V., Che, H., Liu, Z., Wang, S.: Lidardm: Generative lidar simulation in a generated world (2024), https://arxiv.org/abs/2404.02903