

AutoScape: Geometry-Consistent Long-Horizon Scene Generation

Jiacheng Chen^{*2} Ziyu Jiang^{*†1} Mingfu Liang³ Bingbing Zhuang¹
Jong-Chyi Su¹ Sparsh Garg¹ Ying Wu³ Manmohan Chandraker^{1,4}

¹NEC Labs America ²Simon Fraser University ³Northwestern University ⁴UC San Diego



Figure 1. AutoScape generates long-horizon and 3D-consistent driving scenes from a single input image, producing high-quality videos over a temporal span of 20 seconds. **Click** the images in the right column within Adobe Reader to play three example videos.

Abstract

This paper proposes AutoScape, a long-horizon driving scene generation framework. At its core is a novel RGB-D diffusion model that iteratively generates sparse, geometrically consistent keyframes, serving as reliable anchors for the scene’s appearance and geometry. To maintain long-range geometric consistency, the model 1) jointly handles image and depth in a shared latent space, 2) explicitly conditions on the existing scene geometry (i.e., rendered point clouds) from previously generated keyframes, and 3) steers the sampling process with a warp-consistent guidance. Given high-quality RGB-D keyframes, a video diffusion model then interpolates between them to produce dense and coherent video frames. AutoScape generates realistic and geometrically consistent driving videos of over 20 seconds, improving the long-horizon FID and FVD scores over the prior state-of-the-art by 48.6% and 43.0%, respectively. Project page: <https://auto-scape.github.io>.

1. Introduction

Recent advances in high-quality video generation are rapidly transforming various applications, ranging from robotics to mixed reality, where the synthesis of realistic visual data is

crucial. A particular example is autonomous driving, where the photorealistic generation of driving videos plays a crucial role in simulation and verification. But despite promising prospects, generating 3D driving scenes that remain coherent and consistent over long horizons remains a fundamental challenge. Current generative methods, while achieving impressive photorealism [15, 17, 23, 81, 82, 89], often struggle with maintaining physical realism [17] and suffer from quality degradation during auto-regressive generation [12, 15], making spatiotemporal coherence over long horizons a critical unsolved challenge.

We introduce **AutoScape**, a novel framework that addresses the challenges of long-horizon 3D-consistent driving scene generation by leveraging explicit geometry awareness. Given the observation that degrading geometric consistency is the key bottleneck of long-horizon scene or video generation, we decompose the problem hierarchically into *sparse RGB-D keyframe generation* and *dense video interpolation*. The core idea is to train a powerful RGB-D diffusion model to generate highly consistent keyframes, which serve as reliable anchors for the scene’s global appearance and geometry, robustly handling a large span. Given the reliable anchors, a video diffusion model then interpolates between keyframes, refining rendered point clouds into a coherent video.

Our RGB-D diffusion model generates high-quality keyframes with three key designs: 1) *joint RGB-D mod-*

^{*}Equal Contribution. [†]Project Lead.

eling, which operates on the joint distribution of color and depth for more coherent appearance and geometry, and conducts pre-training on large-scale paired data from diverse sources beyond driving to obtain general RGB-D priors; 2) *explicit geometry conditioning*, where each generated RGB-D keyframe is directly conditioned on the existing scene’s appearance and geometry, in the format of rendered point clouds; 3) *warp consistent guidance*, a classifier guidance style approach that steers the diffusion model’s sampling process toward better geometric alignment with previous keyframes, mitigating the accumulation of errors throughout long-term generation.

Compared to those methods that handle spatial and temporal consistency using only the temporal modules of a video diffusion model [17, 19], our hierarchical approach offers greater robustness in terms of long-horizon generation, since the RGB-D diffusion model first produces sparse yet highly consistent keyframes as global anchors rather than directly generating dense frames. Compared to existing works that also employ explicit 3D modeling and produce keyframes [82], our method demonstrates superior keyframe quality by the joint RGB-D modeling, geometry conditioning, and warp-consistent guidance.

As shown in Figure 1, AutoScape generates long-horizon, 3D-consistent, and high-quality scenes with a video duration of 20 seconds containing 250 frames. Quantitatively, it achieves significant improvements over the previous state-of-the-art method, with reductions of 48.6% and 43.0% in FID and FVD scores, respectively, in terms of long-horizon video generation. To summarize, our contributions are threefold:

- **AutoScape**, a novel framework that jointly generates the appearance and geometry of long-range driving scenes using a hierarchical approach of keyframe generation and interpolation.
- A new RGB-D diffusion model featuring geometry-aware conditioning and guidance to enforce long-range 3D consistency, ensuring both geometric stability and high-fidelity visual quality.
- State-of-the-art quantitative and qualitative results in long-horizon driving scene generation as demonstrated by comprehensive experiments.

2. Related Works

Diffusion Models. Diffusion-based generative models [22, 57, 59] have fueled a surge in generative AI. While the theoretical advancements keep improving the mathematical formulation, sampling speed, and generation quality [27, 36, 39, 58, 60], Variants of diffusion models have extended the early success in image generation [13, 48, 51] to a broad spectrum, including video [3, 4, 72], audio [31, 38], 3D [11, 30, 33, 34, 43, 47, 65, 85], motion synthesis [26, 88], visual editing [5, 9, 45, 69, 73], and more. We employ diffusion models for RGB-D keyframe generation and interpolation

toward long-horizon driving scene generation.

3D Scene Generation. Diffusion models have been widely applied in scene generation. One prominent framework is iterative inpainting, where scenes are progressively expanded from an initial image, like SceneScape [15], Text2Room [23], and WonderJourney [82]. WonderWorld [81] advances this paradigm by enhancing 3D consistency and supporting interactive user control. Another widely adopted paradigm involves generating 360-degree panoramas that can be converted into 3D models [14, 55, 64, 90]. Additional methods focus on producing high-level scene layouts [41, 56, 74] or generating LiDAR point clouds to represent 3D scene structures [49, 87]. In autonomous driving, methods like MagicDrive3D [17] and DriveDreamer4D [89] generate driving videos using diffusion models, which are subsequently transformed into 3D scenes for efficient simulation. However, the temporal span of video diffusion models constrains these approaches in terms of the scene scale.

Street View Generation. Although numerous studies investigated street-view generation with reconstruction systems [10, 62, 66, 75, 76, 79], the recent advances in diffusion models have made generative simulation popular. Diffusion-based approaches have been applied to sensor simulation or data augmentation, leveraging layout conditioning such as HD maps or object bounding boxes to produce realistic urban scenes [16, 24, 37, 53, 63, 68, 70, 77]. MagicDrive [16] and MVPbev [37] employ cross-view attention mechanisms to synthesize multi-camera images. StreetScapes [12] proposes an autoregressive video diffusion model to generate long-range street-view videos, conditioned on 2.5D maps. Vista [19] constructs a driving model capable of producing extended, high-fidelity driving videos with action controls. DriveArena [78] integrates components from previous methods to establish a generative closed-loop simulator. Our work focuses on scene and video generation over substantially longer temporal horizons.

Concurrent Works. Several concurrent efforts explore long-horizon street-view synthesis [18, 42]. InfiniCube [42] constructs an explicit sparse-voxel 3D world to guide a video diffusion model, generating unbounded driving scenes. MagicDrive-V2 [18] introduces an efficient video-diffusion architecture that scales to longer sequences. In contrast, our approach develops a novel RGB-D diffusion model that iteratively produces sparse, geometry-consistent keyframes, which in turn facilitate long-horizon video generation.

3. Preliminary

Diffusion models lay the foundation of our scene generation framework. Diffusion-based generative models [22, 57, 59] have recently emerged as a dominant family of generative models, capable of capturing complex data distributions through iterative denoising processes. The core mechanism involves a pre-defined forward diffusion process

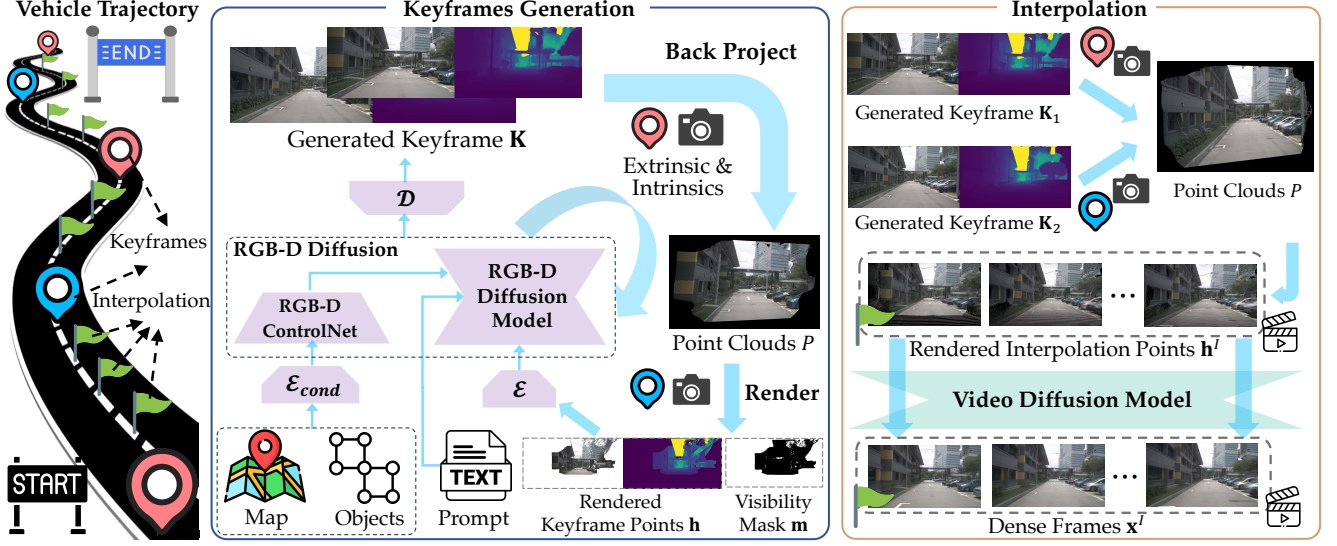


Figure 2. Pipeline of the AutoScope. The vehicle trajectory defines the location of keyframes and interpolation frames, spanning a long-horizon 3D space. The *Keyframes Generation* stage iteratively generates geometrically consistent keyframes with an RGB-D diffusion model as global scene anchors. The *Interpolation* stage then produces dense frames with a video diffusion model. The keyframe viewpoints are indicated by $\odot$ and the interpolation viewpoints are marked by $\triangle$, best viewed in color.

$q(\mathbf{x}_t | \mathbf{x}_{t-1})$ that incrementally adds Gaussian noise to the data over T timesteps, transforming an original data sample $\mathbf{x}_0$ into a noisy $\mathbf{x}_T$, defined by:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

where β_t denotes the variance schedule controlling the noise level at each timestep, and $\mathbf{I}$ is the identity matrix. The *reverse process* aims to recover the original data by learning a parameterized denoising model $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (2)$$

$\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ representing the mean and covariance functions modeled by neural networks with parameters θ . By iteratively applying this process starting from a Gaussian noise $\mathbf{x}_T$, the model generates new data samples that resemble the training data distribution.

Latent Diffusion Models (LDMs) [51] operate in the compressed latent space of a pre-trained autoencoder rather than the raw high-dimensional data space, which enhances computational efficiency without compromising generative performance. Our method is based on the LDM formulation.

4. AutoScope

4.1. Framework Overview

This paper focuses on generating long-horizon, high-quality, and 3D-consistent driving scenes. Although recent advancements in general video generation techniques [3, 72] have made promising progress in producing driving videos [16,

19], ensuring 3D consistency across hundreds of frames (e.g., over 20 seconds) remains hard. Long-range temporal and geometric consistency are the central challenges.

AutoScope is a two-stage scene generation framework aiming for robust long-term coherency and stability (Figure 2). In the *keyframe generation* stage, a RGB-D diffusion model jointly generates keyframes and the corresponding point clouds to anchor the scene’s global appearance and geometry. In the *interpolation* stage, dense frames are first rendered from the consecutive RGB-D keyframes and then refined into coherent images using a video diffusion model. The explicit geometry modeling makes the first stage produce consistent yet sparse keyframes, which then serve as reliable conditions for the interpolation stage.

Keyframe Generation Process. The keyframe generation process, illustrated in Figure 2 (left), generates keyframes iteratively along specified sparse viewpoints. Each iteration comprises three steps: back-projection, rendering, and diffusion model generation. In the first iteration, we begin with a real input image or a generated RGB-D image, back-projecting the image into 3D space as point clouds $\mathcal{P}$ with camera parameters. The back projection is defined as:

$$\mathcal{P} = \mathbf{B}(\mathcal{X}_{\text{rgb}}, \mathcal{X}_{\text{depth}}, \mathcal{C}), \quad (3)$$

where $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^N$ denotes the set of 3D point clouds obtained from the existing N keyframes. $\mathcal{X}_{\text{rgb}} = \{\mathbf{x}_{\text{rgb},i}\}_{i=1}^N$ and $\mathcal{X}_{\text{depth}} = \{\mathbf{x}_{\text{depth},i}\}_{i=1}^N$ represent the collections of RGB images and depth images of the keyframes, respectively. $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^N$ is the set of camera parameters (including intrinsics and extrinsics) corresponding to each keyframe.

$\mathbf{B}(\cdot)$ is the back-projection function that reconstructs the 3D point clouds $\mathbf{P}_i$ from the RGB and depth images using the camera parameters c_i for each keyframe i . Subsequently, the rendered keyframe points are produced by projecting the point clouds onto the image plane of the next keyframe:

$$\mathbf{h}, \mathbf{m} = \mathbf{R}(\mathcal{P}, c), \quad (4)$$

where $\mathbf{h}$ is the rendered keyframe points, essentially a coarse image with noise and holes. $\mathbf{m}$ is the corresponding visibility mask indicating the presence of projected points. $\mathbf{R}(\cdot)$ is the rendering function that projects the 3D point clouds onto the image plane defined by the target camera parameters c .

The rendered keyframe points $\mathbf{h}$ and the visibility mask serve as the conditioning input for the RGB-D diffusion model, along with the map, object boxes, and prompt conditions. The generated RGB-D keyframe $\mathbf{K}$ would then contribute to the next iteration by adding its back-projection into the existing point clouds. The auto-regressive process runs in reverse along the trajectory, starting from the end and iteratively moving to the next nearest keyframe viewpoint until it reaches the start. § 4.2 presents the details of this RGB-D diffusion model.

Recent works, such as WonderJourney [81, 82], have also explored the use of keyframes. However, these methods primarily rely on pretrained image inpainting models for keyframe generation. We propose a novel conditional RGB-D diffusion model, which offers key advantages in terms of **generalizability** and **geometry awareness**. Specifically, we introduce a two-stage training framework that enables the diffusion model to be pre-trained on large-scale RGB-D data (millions of images), thereby improving its **generalizability**. Moreover, previous methods predict depth solely from the generated RGB frame, which restricts the ability of different frames to access details from the previously generated geometry, often resulting in inconsistent depth maps. To overcome this limitation, we make the diffusion model conditioned on both appearance and geometry, thus generating more coherent keyframes. This design significantly enhances the model’s **geometry awareness** and ensures better consistency with existing scene geometry.

Warp Consistent Guidance. While the explicit geometry conditioning improves the long-term consistency, we still observe that the generated content of the RGB-D diffusion model occasionally misaligns with the rendered keyframe points from previous keyframes. To further improve the consistency at test time, we propose a warp consistent guidance mechanism, steering the sampling process toward better 3D alignment during inference. More details are in § 4.3

Interpolation. The interpolation process generates dense frames by interpolating between two consecutive keyframes. The interpolation process is conditioned on the rendered 3D point clouds derived from the keyframes and utilizes an off-the-shelf video diffusion model from ViewCrafter [83].

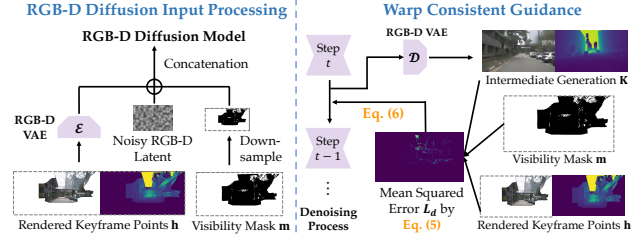


Figure 3. (Left) Input processing of our RGB-D diffusion model. (Right) The warp-consistent guidance for steering the sampling process toward better geometric consistency.

Existing approaches [81, 82] rely on pre-trained image generation models to interpolate sparse keyframes, neglecting the global consistency of the interpolated frames. Given high-quality keyframes generation, a point-cloud-conditioned video diffusion model can produce highly coherent interpolation along the rendering trajectory.

In the rest of the section, § 4.2, § 4.3, and § 4.4 elaborate on the RGB-D diffusion model, the warp consistent guidance mechanism, and the interpolation video diffusion models.

4.2. RGB-D Diffusion for Keyframe Generation

The core of AutoScape is a diffusion model that auto-regressively generates RGB-D keyframes. It incorporates robust geometric priors by explicitly modeling the depth and training on a large dataset with curated depth information. The model progressively builds upon the existing scene’s geometry by conditioning on the rendered keyframe points. This design enhances both the appearance and geometric coherency between the new and previously generated keyframes, thereby maintaining quality and consistency across a long generation horizon.

Backbone. Our RGB-D diffusion model is based on the Latent Diffusion Model (LDM) [51]. It comprises a Variational Autoencoder (VAE) [28] that compresses inputs into a latent space, where the denoising U-Net [52] is trained to revert the forward process. To extend the model for RGB-D conditioning and generation, we first extend the VAE to jointly model image and depth, encoding and decoding the RGB-D data to or from the latent space, respectively. Full details of the RGB-D VAE are in § B of the Appendix.

Conditioning. To integrate the Rendered Keyframe Points $\mathbf{h}$ as a condition, as shown in Figure 3 (left), it is first encoded by the RGB-D VAE, serving as an additional conditioning input to the model. The visibility mask $\mathbf{m}$ is down-sampled to match the spatial resolution of the latent space. The noisy latent, the mask, and the RGB-D latent code of $\mathbf{h}$ are concatenated along the channel dimension and fed into the U-Net. To accommodate the additional channels, the U-Net architecture is extended by adding five extra input channels. The initial convolution layers for processing these new channels are zero-initialized. Following Stable Diffusion [51], text prompts are injected through the cross-attention module. We

also incorporate HD maps and object boxes through ControlNet [86]; details are discussed in § A.

Training Pipeline. The training of the RGB-D diffusion model comprises two stages: RGB-D pre-training and rendering-conditioned fine-tuning. The RGB-D pre-training stage performs the RGB-D inpainting task on large-scale, curated image data. The depth used for training is pseudo-labeled with Metric3D [25], an off-the-shelf monocular metric depth estimator. Note that the additional conditions are simulated by masking the data with synthetic masks, rather than being derived from a keyframe, which is a key for large-scale training on diverse data sources. In the rendering-conditioned fine-tuning stage, the rendered keyframe points and visibility mask are derived from the last keyframe, while the map and bounding box conditions are also incorporated. The model is fine-tuned on driving-specific datasets. More training details are provided in § 5.1.

4.3. Warp Consistent Guidance

Although diffusion models conditioned on rendered keyframe points (*i.e.*, coarse image) share similarities with traditional image inpainting, we observe that the generated content sometimes exhibits pronounced appearance and geometry inconsistencies in the overlapping regions. One potential cause is the noisy training data, where the depths of two consecutive keyframes do not align perfectly with each other. The inconsistency adversely affects 3D consistency, leading to increasingly noticeable shifts in appearance and scene geometry throughout the iterative generation process.

To mitigate this, we propose *Warp Consistent Guidance* (WCG) to steer the sampling process of the diffusion model toward better geometric consistency. The idea is to introduce a projection consistency loss to quantify the discrepancy between rendered keyframe points and RGB-D generation, as illustrated in Figure 3 (right). The loss then adjusts the sampling process through classifier guidance. Concretely, the loss is defined as the masked Mean Squared Error (MSE) between the predicted RGB-D frame $\mathbf{x}$ and the rendered keyframe points input $\mathbf{h}$:

$$\mathcal{L}_d(\mathbf{x}, \mathbf{h}; \mathbf{m}) = \frac{\sum_i \mathbf{m}_i (\mathbf{x}_i - \mathbf{h}_i)^2}{\sum_i \mathbf{m}_i}. \quad (5)$$

where i is the pixel index, and we slightly abuse the subscript of $\mathbf{x}$ here. $\mathbf{m}_i \in \{0, 1\}$ is the i -th pixel of the overlap mask $\mathbf{m}$. $\mathbf{m}_i = 1$ means the pixel is visible in both the target keyframes and the previously generated keyframes, and $\mathbf{m}_i = 0$ otherwise. We mask out 5% pixels with the largest loss for better robustness against noise.

The gradient of $\mathcal{L}_d$ then guides the generation towards latent regions that are more geometrically consistent with existing keyframes. Formally, at timestep t , the sampling process is modified by adjusting the original score estimate

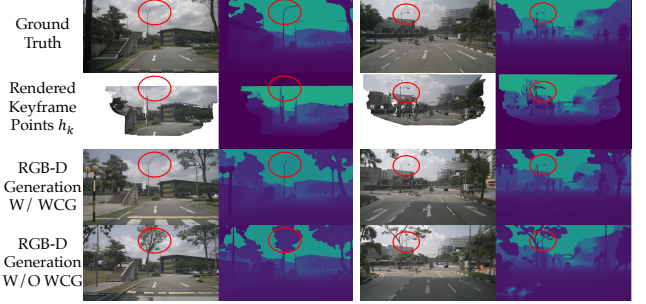


Figure 4. Qualitative comparison to show the effectiveness of our warp consistent guidance (WCG) strategy for better consistency.

$\mathbf{s}_\theta(\mathbf{x}_t, t)$ with $\nabla_{\mathbf{x}_t} \mathcal{L}_d$. The adjusted score is defined as:

$$\tilde{\mathbf{s}}_\theta(\mathbf{x}_t, t) = \mathbf{s}_\theta(\mathbf{x}_t, t) + w \nabla_{\mathbf{x}_t} \mathcal{L}_d(\mathbf{x}_t, \mathbf{h}; \mathbf{m}), \quad (6)$$

$\tilde{\mathbf{s}}_\theta(\mathbf{x}_t, t)$ is the adjusted score used at each sampling step, and w is the scale that controls the strength of WCG. Figure 4 demonstrates that WCG significantly improves the consistency between the rendered keyframe points and the newly generated keyframe, which eventually enhances the quality and stability of the long-horizon generation.

4.4. Video Diffusion Model for Interpolation

After generating the sparse RGB-D keyframes along the trajectory in the 3D scene, the interpolation stage connects two consecutive keyframes with dense frames, as illustrated in Figure 2 (right). Compared to an image inpainting model, employing a video diffusion model enables smoother interpolation due to its strong temporal priors. Furthermore, the video diffusion model can be anchored on our high-quality RGB-D keyframes. The rendering results provide effective geometric cues, making it easier to produce coherent interpolation. The interpolation process is defined by:

$$\{\mathbf{x}_t^I\}_{t=1}^T = G(\{\mathbf{z}_t\}_{t=1}^T; \{\mathbf{h}_t^I\}_{t=1}^T, \mathbf{K}_1, \mathbf{K}_2). \quad (7)$$

G is the video diffusion model, and we use the off-the-shelf point-cloud-conditioned model from ViewCrafter [83]. T is the number of interpolation frames between two keyframes. Each $\mathbf{z}_t$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. $\{\mathbf{x}_t^I\}_{t=1}^T$ are the output frames and $\{\mathbf{h}_t^I\}_{t=1}^T$ are the corresponding rendered keyframe points. Unlike the rendered results $\mathbf{h}$ used in keyframe generation, $\mathbf{h}^I$ only contains the RGB information without depth to fit the pre-trained model’s input specification. $\mathbf{K}_1$ and $\mathbf{K}_2$ are the two consecutive RGB-D keyframes.

5. Experiment

Baselines. We compare AutoScape with Vista [19] and WonderJourney [82], two competitive methods for generating long-horizon videos. Vista is the state-of-the-art in driving video generation. WonderJourney achieves long-horizon

Table 1. Comparison of Methods in Terms of FID and FVD at Different Time Splits. WonderJourney[†] indicated WonderJourney adapted for driving scene. Overall indicates the time split of 0-20s. For both FID and FVD, lower is better, denoted by ↓. The top-performing methods are highlighted in **bold**. The proposed method advances the previous State-of-the-Art method Vista [19] by reducing FID from 68.3 to 35.1 and FVD from 629.8 to 359.0, corresponding to a significant improvement margin of 48.6% and 43.0%, respectively.

Method	0–5s		5–10s		10–15s		15–20s		Overall	
	FID↓	FVD↓	FID↓	FVD↓	FID↓	FVD↓	FID↓	FVD↓	FID↓	FVD↓
WonderJourney [82]	93.0	977.2	157.4	1651.7	172.7	1716.7	172.5	1737.8	127.5	1017.4
WonderJourney [†] [82]	49.8	661.8	111.7	1551.5	114.1	1730.1	99.0	1756.7	73.7	939.6
Vista [19]	37.2	436.4	72.4	967.0	124.6	1329.1	157.9	1614.5	68.3	629.8
AutoScape (Ours)	34.3	385.9	48.8	526.3	54.0	579.5	56.8	657.4	35.1	359.0

generation for general scenes. We adapt WonderJourney by finetuning its diffusion model on driving data and performing the generation with vehicle trajectories from nuScenes (denoted as WonderJourney[†]) for fair comparisons.

Datasets and Evaluation Metrics. We evaluate all methods on the nuScenes validation set (with 150 videos) using Fréchet Inception Distance (FID) [20] and Fréchet Video Distance (FVD) [67] while extending the evaluation to focus more on **long horizon**. Each model generates a long sequence of frames from a single input image, and the generation can extend up to 20 seconds. To better understand the performance change over time, we compute FID and FVD scores over consecutive 5-second segments.

In the rest of this section, §5.1 introduces the primary implementation details of AutoScape, then §5.2, §5.3 and §5.4 present the evaluation results and analyses. We refer to the supplementary for more details and results.

5.1. Implementation Details

RGB-D Pre-training. RGB-D pre-training scales the diffusion model on large-scale RGB-D datasets to learn robust geometry priors. While there are many existing high-quality image datasets [7, 54], depth data is scarce. To scale up the training, we generate depth with a monocular metric depth predictor [25, 80]. In practice, we collect RGB images from nuScene (training split) [6], Argoverse2 (training split) [71], and SA1B [29] and prepare the depth data, forming a dataset of 13 million diverse images. We use the ground-truth intrinsics for the depth predictor [25] on nuScene and Argoverse2, while predicting the intrinsics with WildCamera [91] on SA1B. We also generate text pseudo-labels with a Vision-Language Model [35] for pretraining. We initialize the Diffusion Unet with the pre-trained Unet of SD-Inpainting-V2.0 [51]. We train the model with text-conditioned RGB-D in-painting to preserve the text controllability and inpainting ability of the base model. The inpainting masks are randomly generated to resemble the visibility mask from projection. We use 32 A100 GPUs for RGB-D pre-training, training for 50k iterations over 2 days. The batch size is 1024 and the learning rate is 1e-4.

Rendering Conditioned Training. For rendering conditioned training, we employ a training strategy mimicking the iterative keyframes generation process. Specifically, each training sample is generated via sampling a pair of frames from the same video sequence with a gap range from 5 to 60 frames. Assigning one of them to be the condition frame and the other as the target, we then project the condition frame to the target frame utilizing the depth and cameras. The projection then serves as the rendered keyframe points conditioning for the target frame. The model is also conditioned with HD maps and object boxes. The training only uses nuScenes, and we prepare 500 samples per scene, resulting in a dataset with 350k samples. The training is conducted on 8 A6000 GPUs over 2 days, with a batch size of 512 for 20k iterations, using a learning rate of 1e-4.

Data Filtering. The rendered keyframe points are sometimes inconsistent with the target image due to noisy depth, dynamic objects, and occlusions. This can impair the 3D consistency of the iterative generation. To alleviate this, we use the warp consistent loss $\mathcal{L}_d(x, h; m)$ to measure the consistency between two frames and filter out the most inconsistent samples. In practice, we filter out 20% of the 350k samples and train only with the remaining 280k.

Video Diffusion Model Training Setting. Given the high-quality RGB-D keyframes, the pre-trained model from ViewCrafter [83] can directly produce promising interpolation results. We therefore use the point-cloud-conditioned model without fine-tuning. Several concurrent methods [2, 44, 50] are also potentially applicable to this stage, and we leave systematic comparisons for future work.

Viewpoints Selection for Keyframes. Selecting an optimal spacing for keyframes is essential. On the one hand, overly dense keyframes result in inefficient modeling of long-range geometric dependencies. On the other hand, if the keyframes are too sparse, the interpolation stage could fail. We designate the first keyframe as one endpoint of the trajectory, then traverse the trajectory to identify the subsequent keyframes. The first viewpoint with either the distance or the view angle difference from the previous keyframe exceeding β or γ , respectively, is selected as the next keyframe. In practice.

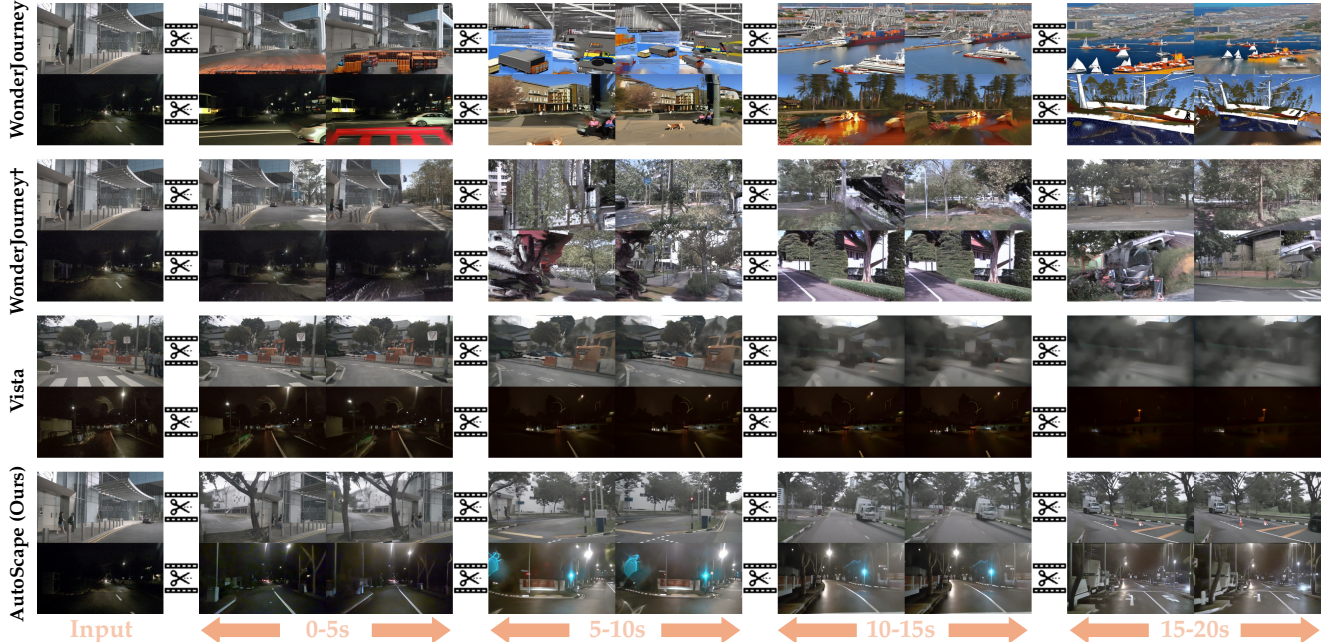


Figure 5. Qualitative comparisons. WonderJourney[†] represents WonderJourney adapted for driving scenes. We sample frames from different time splits to demonstrate both short and long-term performance. AutoScape maintains 3D consistency over significant view changes and extends toward a long horizon. Vista generates the video from the start of the trajectory to the end, while others generate in the reverse order.

We set $\beta = 10m$ and $\gamma = 20^\circ$.

5.2. Main Results

Quantitative Comparison. Table 1 shows that AutoScape achieves the best quantitative performance across all time splits. For overall video quality, it improves Vista from [68.3, 629.8] to [35.1, 359.0] in terms of [FID, FVD], respectively. This margin is even larger in terms of the long-horizon generation in time split of 15-20s, from [157.9, 1614.5] to [56.8, 657.4] in terms of [FID, FVD], corresponding to a significant improvement of 64% and 59.3%. To evaluate generalizability, we further assess our method on the Argoverse2 dataset [71] without fine-tuning. While Vista achieved [FID, FVD] scores of [80.4, 614.2], AutoScape significantly improves the scores to [49.2, 317.9], demonstrating the robust generalizability achieved through the RGB-D pre-training.

Qualitative Comparison. As visualized in Figure 5, while WonderJourney produces reasonable images across all time splits, it struggles to generate photorealistic driving scenes. After adapting it to driving data, WonderJourney[†] shows better photorealism. However, both versions exhibit a gradual loss of context and deviate from the input image. For instance, in the video clip spanning 5–10s, both versions alter the scene from night to day. Additionally, WonderJourney depends heavily on carefully designed camera trajectories, and the results deteriorate with real-world vehicle trajectories. Vista, on the other hand, can generate high-quality videos for short clips (e.g., 0–5 seconds), but the performance rapidly drops for more extended sequences. In

Table 2. Ablation study on the design components of AutoScape.

Methods	FID↓	FVD↓
AutoScape (Ours)	35.1	359.0
– RGB-D Pre-training	47.6	650.0
– Data Filtering	43.5	463.0
– Warp Consistent Guidance	38.5	380.2
– Depth Generation	39.2	511.4

contrast, AutoScape consistently produces 3D-consistent, long-horizon, high-quality videos across all temporal splits, effectively handling significant dynamics of real-world trajectories and generalizing well to challenging conditions.

User Study. To thoroughly evaluate long-sequence 3D consistency, we conducted a user study in which participants selected the video exhibiting the best 3D consistency performance over a long sequence. From 22 valid responses, our method was preferred in 88.39% of the cases against Vista and both variants of WonderJourney.

5.3. Ablation Study

We investigate the effectiveness of different components of AutoScape in Table 2 and analyze the results below.

RGB-D Pretraining. For this ablation, instead of employing the two-stage training pipeline, we initialize the model with the Stable Diffusion model and finetune it with only the Rendering Conditioned Training stage. Removing the pre-training stage significantly impacts performance, increasing the FID and FVD from 35.1 and 359.0 to 47.6 and 650.0,

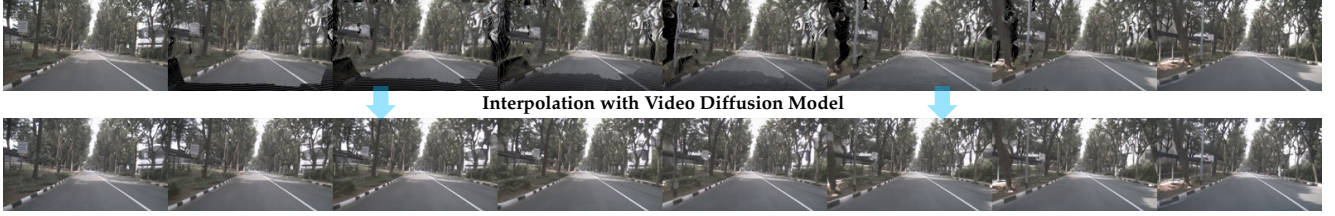


Figure 6. Illustration of the interpolation process. (Top) The intermediate frames between two RGB-D keyframes with the rendered interpolation points. (Bottom) The corresponding interpolation results from the video diffusion model.



Figure 7. Generalization to various weather conditions, rare objects, and uncommon scene types.

respectively, highlighting the effectiveness of large-scale RGB-D pre-training.

Data Filtering. As introduced earlier, we filter out 20% most inconsistent samples for training. Training with noisy data results in significant performance degradation of 8.4 and 104 in FID and FVD, respectively. This indicates that the inconsistency between the rendered keyframe points and the generated content has a significant impact on the consistency.

Warp Consistent Guidance. Warp consistent guidance can significantly boost the FID and FVD of the generated video by further enhancing the 3D consistency. Removing it would result in a performance drop from 35.1 to 38.5 for FID and from 359.0 to 380.2 for FVD, indicating the effectiveness of the proposed Warp Consistent Guidance.

Joint Depth Generation. We then study the necessity of jointly generating depth with the diffusion model. For this ablation study, we replace the RGBD Diffusion model with the standard RGB Diffusion model. The depth is predicted using a monocular metric depth prediction model [25] based on the generated RGB image. As shown in Table 2. Removing depth generation decreases the FID and FVD from 35.1 and 359.0 to 39.2 and 511.4, respectively. The clear degradation of FVD indicates that the model without depth generation lacks an understanding of geometry and leads to worse 3D consistency.

5.4. More Results and Analyses

Interpolation Process. Figure 6 illustrates the process of interpolation generation with the intermediate rendering condition. The projection of point clouds depicts the appearance of the global frame, ensuring the 3D consistency of generated content. And the video diffusion model complements the missing holes and produces a high-quality video clip. More examples can be found in the supplementary material.

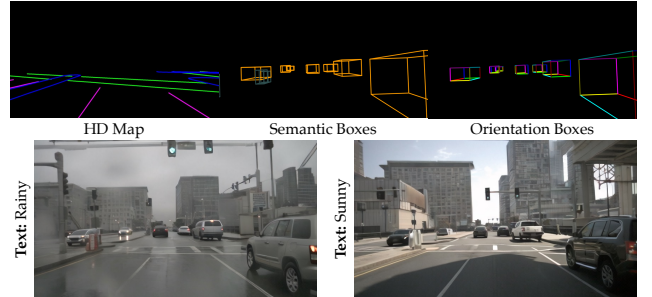


Figure 8. AutoScape’s generation controllability with texts (for weather), HD maps, and object bounding boxes.

Generalization to Corner Case. Figure 6 illustrates how AutoScape handles out-of-distribution (OOD) corner cases, including rare weather conditions (e.g., rainy), uncommon vehicle types (e.g., ambulances), and uncommon environments (e.g., rural landscapes). Figure 5 also covers the generation of night-time scenes.

Controllability. Figure 8 demonstrates the fine-grained controllability of our method through texts, object bounding boxes, and HD maps. The flexible control enables the generation of highly customized, long-horizon scenes.

6. Conclusion

This paper introduced AutoScape, a hierarchical framework for long-horizon, 3D-consistent driving scene generation. The core is a novel RGB-D diffusion model that generates geometrically-consistent keyframes using joint RGB-D modeling, explicit geometry conditioning, and warp-consistent guidance. By anchoring a video interpolation stage with high-quality keyframes, AutoScape successfully mitigates geometric drift, generating realistic and coherent driving scene videos that maintain consistency for over 20 seconds.

References

- [1] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025. 2
- [2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 6
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3
- [4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023. 2
- [6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020. 6
- [7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021. 6
- [8] Jiacheng Chen, Yuefan Wu, Jiaqi Tan, Hang Ma, and Yasutaka Furukawa. Maptracker: Tracking with strided memory fusion for consistent vector hd mapping. In *European Conference on Computer Vision (ECCV)*, 2024. 1
- [9] Jiacheng Chen, Ramin Mehran, Xuhui Jia, Saining Xie, and Sanghyun Woo. Blenderfusion: 3d-grounded visual editing and generative compositing. *arXiv preprint arXiv:2506.17450*, 2025. 2
- [10] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024. 2
- [11] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [12] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 1, 2
- [13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems (NeurIPS)*, 2021. 2
- [14] Chuan Fang, Yuan Dong, Kunming Luo, Xiaotao Hu, Rakesh Shrestha, and Ping Tan. Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. *arXiv preprint arXiv:2310.03602*, 2023. 2
- [15] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2
- [16] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 2, 3, 1
- [17] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024. 1, 2
- [18] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024. 2
- [19] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 2, 3, 5, 6
- [20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems (NeurIPS)*, 2017. 6
- [21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 2020. 2
- [23] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2
- [24] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2
- [25] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 5, 6, 8

- [26] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [27] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems (NeurIPS)*, 2022. 2
- [28] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4
- [29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6
- [30] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [31] Zhifeng Kong, Wei Ping, Jiayi Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 2
- [32] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapi: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, 2022. 1
- [33] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024. 2
- [34] Chuang Lin, Bingbing Zhuang, Shanlin Sun, Ziyu Jiang, Jianfei Cai, and Manmohan Chandraker. Drive-1-to-3: Enriching diffusion priors for novel view synthesis of real vehicles. *arXiv preprint arXiv:2412.14494*, 2024. 2
- [35] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6
- [36] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2
- [37] Buyu Liu, Kai Wang, Yansong Liu, Jun Bao, Tingting Han, and Jun Yu. Mvpbev: Multi-view perspective image generation from bev with test-time controllability and generalizability. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. 2
- [38] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023. 2
- [39] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2
- [40] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1
- [41] Jack Lu, Kelvin Wong, Chris Zhang, Simon Suo, and Raquel Urtasun. Scenecontrol: Diffusion for controllable traffic scene generation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 2
- [42] Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. *arXiv preprint arXiv:2412.03934*, 2024. 2
- [43] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [44] YU Mark, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectoryrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 6
- [45] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2
- [46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems (NeurIPS)*, 2019. 1
- [47] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2
- [48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2
- [49] Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with lidar diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [50] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 6
- [51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022. 2, 3, 4, 6, 1
- [52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference*,

Munich, Germany, October 5-9, 2015, proceedings, part III 18, 2015. 4

- [53] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025. 2
- [54] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6
- [55] Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, et al. Controlroom3d: Room generation using semantic proxy rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [56] Mohammad Amin Shabani, Sepidehsadat Hosseini, and Yasutaka Furukawa. Housediffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning (ICML)*. PMLR, 2015. 2
- [58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2
- [60] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 2
- [61] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023. 1
- [62] Shanlin Sun, Bingbing Zhuang, Ziyu Jiang, Buyu Liu, Xiaohui Xie, and Manmohan Chandraker. Lidarf: Delving into lidar for neural radiance field on street scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [63] Alexander Szwedlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird’s-eye view layout. *IEEE Robotics and Automation Letters*, 2024. 2
- [64] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *ArXiv*, abs/2307.01097, 2023. 2
- [65] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024. 2
- [66] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14895–14904, 2024. 2
- [67] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6
- [68] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jia-gang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 2, 1
- [69] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhui Chen. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2024. 2
- [70] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2
- [71] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 6, 7
- [72] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3
- [73] Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey Allen, and Thomas Kipf. Neural assets: 3d-aware multi-object scene synthesis with image diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2
- [74] Chejian Xu, Ding Zhao, Alberto Sangiovanni-Vincentelli, and Bo Li. Diffscene: Diffusion-based safety-critical scenario generation for autonomous vehicles. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023. 2
- [75] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pages 156–173. Springer, 2024. 2
- [76] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023. 2

- [77] Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint arXiv:2308.01661*, 2023. 2
- [78] Xuemeng Yang, Licheng Wen, Yukai Ma, Jianbiao Mei, Xin Li, Tiantian Wei, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. *arXiv preprint arXiv:2408.00415*, 2024. 2
- [79] Ze Yang, Yun Chen, Jingkan Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 2
- [80] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6
- [81] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024. 1, 2, 4
- [82] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 4, 5, 6
- [83] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 4, 5, 6
- [84] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 1
- [85] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *ArXiv*, abs/2210.06978, 2022. 2
- [86] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5, 1
- [87] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*, 2023. 2
- [88] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2024. 2
- [89] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. *arXiv preprint arXiv:2410.13571*, 2024. 1, 2
- [90] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suyu You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024. 2
- [91] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: in-the-wild monocular camera calibration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 6

AutoScape: Geometry-Consistent Long-Horizon Scene Generation

Supplementary Material

In the supplementary material, we provide additional content that could not be included in the main paper due to page and format constraints. The supplementary material is organized as follows:

- In § A presents the remaining implementation details.
- In § B presents the architectural and training details of the RGB-D VAE of AutoScape.
- In § C provides additional experimental results.

A. Remaining Implementation Details

This section presents the remaining implementation details that are not covered in the main paper due to space limitations. The proposed method is implemented with PyTorch [46] and the Diffuser library.

Optimization Settings. For both RGB-D pretraining and rendering-conditioned training, we utilize the AdamW optimizer to facilitate optimization. The learning rate (lr) and weight decay (wd) are set to 1×10^{-4} and 1×10^{-2} , respectively, with a learning rate warmup applied over the first 3000 iterations. Gradient clipping with a maximum norm of 1 is implemented to enhance training stability. Additionally, both training and inference are conducted using bfloat16 (brain floating-point 16-bit) precision to ensure computational efficiency and optimization effectiveness.

HD Map and Bbox Condition. To enable more flexible controllability, we augment our RGB-D diffusion model with a ControlNet [86] branch to encode HD maps and object bounding boxes. Figure 9 provides a visualization of these conditioning inputs. Specifically, for the map condition, we extract the layers (*i.e.*, lane boundary, lane divider, and pedestrian crossings from the vector HD maps [8, 32, 84] and then project them onto the image plane. To specify the location and orientation of objects precisely, we utilize two types of box control images: semantic box control and orientation box control. Both box controls are derived by projecting 3D bounding boxes onto the image plane with the camera parameters. For the semantic box control, different colors are used to distinguish vehicles, pedestrians, roadblocks, etc. For the orientation box control, the orientation of each vehicle is indicated by assigning unique colors to each edge of the box. Figure 8 in the main paper demonstrates the controlled generation through these protocols. Note that our conditioning strategies for HD Maps and objects are different from those in MagicDrive [16] or DriverDreamer [68].

Training with ControlNet. ControlNet is only incorporated during the rendering-conditioned training stage, as the HD maps and object boxes conditions are not available for the RGB-D pre-training stage, where we use datasets

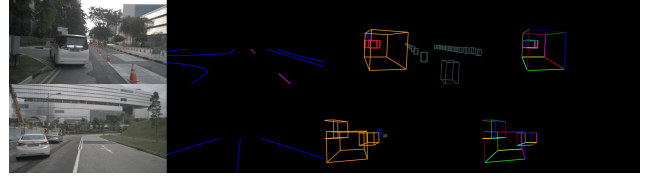


Figure 9. The control signals with the corresponding images. From left to right are ground-truth RGB images, projected maps, semantic box control, and orientation box control.

beyond driving. The ControlNet is initialized using the U-Net model from the RGB-D pretraining stage, following those outlined in the original ControlNet [86]. During the rendering-conditioned training stage, we fine-tune both ControlNet and U-Net to facilitate convergence.

Inference Settings. For diffusion model inference, we utilize DPM-Solver [40] with 50 steps. Additionally, classifier-free guidance [21] is employed to enhance the quality of conditioned generation, using a guidance strength of 7.5 in accordance with the default settings of the diffuser library.

B. Details of the RGB-D VAE

Similar to LDM3D [61], we modify the VAE to support depth encoding and decoding to accommodate depth generation, while preserving the latent code shape. Specifically, we first normalize the depth to 0-1, with a maximum depth of 300 meters, to align with the scale of the RGB channels. Then, the normalized depth (1 channel) is concatenated with RGB (3 channels) to create a 4-channel RGB-D input for the VAE. Architecturally, we extend the first and last convolutions in both the encoder and decoder to accommodate this 4-channel input and output, ensuring compatibility with RGB-D data. As the default 8-bit choice for RGB channel leads to significant precision loss for depth channel [61], we employ 16-bit precision for RGB-D inputs and outputs to retain depth details accurately. Since the latent feature shape remains unchanged, we apply the existing U-Net architecture directly for latent diffusion.

The RGB-D VAE is initialized with the pretrained RGB VAE from Stable Diffusion models [51]. The added parameters are initialized to zero to preserve pretrained knowledge. The optimization target is defined as

$$\begin{aligned} \mathcal{L}_{\text{VAE}} = & \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [-\log p_{\theta}(\mathbf{x}_{\text{rgb}} | \mathbf{z})] \\ & + \lambda_{\text{depth}} \cdot \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [-\log p_{\theta}(\mathbf{x}_{\text{depth}} | \mathbf{z})] \quad (8) \\ & + D_{\text{KL}}(q_{\phi}(\mathbf{z} | \mathbf{x}) || p(\mathbf{z})) \end{aligned}$$

where $\mathbf{x}_{\text{rgb}}$ represents the RGB image data. $\mathbf{x}_{\text{depth}}$ represents

the depth map data. $\mathbf{x}$ is the combination of $\mathbf{x}_{\text{rgb}}$ and $\mathbf{x}_{\text{depth}}$. $q_\phi(\mathbf{z} \mid \mathbf{x}_{\text{rgb}}, \mathbf{x}_{\text{depth}})$ is the encoder network with parameters ϕ , encoding both RGB and depth inputs. $p_\theta(\mathbf{x}_{\text{rgb}} \mid \mathbf{z})$ and $p_\theta(\mathbf{x}_{\text{depth}} \mid \mathbf{z})$ are the decoder networks reconstructing RGB images and depth maps from the latent variable $\mathbf{z}$. D_{KL} is the Kullback-Leibler divergence between the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ and the prior $p(\mathbf{z})$.

The first and second term, $\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} [-\log p_\theta(\mathbf{x}_{\text{rgb}} \mid \mathbf{z})]$ and $\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})} [-\log p_\theta(\mathbf{x}_{\text{depth}} \mid \mathbf{z})]$, minimize the reconstruction errors for the RGB images and depth maps, respectively. The third term $D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z}))$, regularizes the latent space by enforcing alignment with a predefined prior distribution, thereby promoting smoothness and continuity in the latent space $\mathbf{z}$.

Given that depth maps tend to contain less high-frequency information than RGB images due to the inherently smooth nature of geometric data, the reconstruction loss for depth is generally smaller than for RGB. To address this imbalance, we introduce a weighting factor, λ_{depth} , to amplify the depth reconstruction loss. In practice, we set $\lambda_{\text{depth}} = 10$.

To train the RGB-D diffusion model, we implement a two-stage training strategy, as outlined in § 5.1.

C. Additional Experimental Results

More baseline results. To further evaluate the quality of keyframes generated by the proposed AutoScape in comparison to the baselines, we apply ViewCrafter to interpolate the keyframes produced by WonderJourney[†]. This results in FID and FVD scores of 59.1 and 858.9, respectively, which are significantly higher than those achieved by AutoScape (35.1 and 359.0). These findings highlight the superior visual quality of the keyframes generated by our method.

Compare with single-stage video diffusion models. To further assess the performance of our proposed two-stage method against the state-of-the-art one-stage approach, we fine-tune COSMOS-Transfer [1] with the HD map from nuScenes and perform autoregressive generation to produce long videos. COSMOS-Transfer achieved an FID of 44.2 and an FVD of 436.1, whereas our method attained 35.1 and 359.0, respectively, demonstrating its clear superiority.