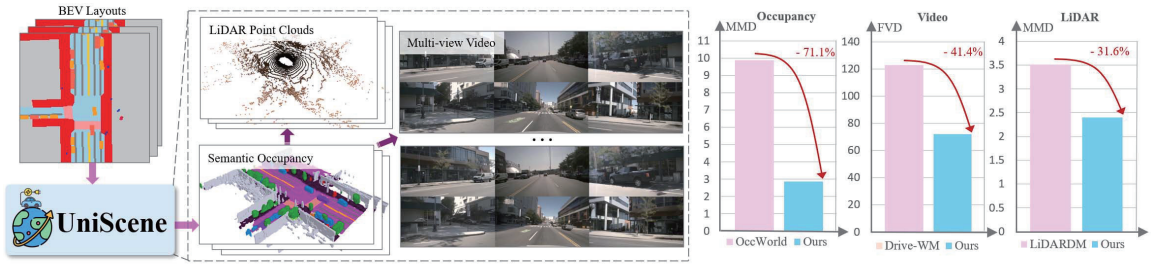
UniScene: Unified Occupancy-centric Driving Scene Generation

Bohan Li[1,2*], Jiazhe Guo[3*], Hongsi Liu[2*], Yingshuang Zou[3*], Yikang Ding[4*], Xiwu Chen[5], Hu Zhu[2], Feiyang Tan[5], Chi Zhang[5], Tiancai Wang[4], Shuchang Zhou[4], Li Zhang[6], Xiaojuan Qi[7], Hao Zhao[3], Mu Yang[4], Wenjun Zeng[2], Xin Jin[2†]

[1]Shanghai Jiao Tong University, [2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]Tsinghua University, [4]MEGVII Technology, [5]Mach Drive, [6]Fudan University, [7]University of Hong Kong



(a) Overview of the proposed UniScene (b) Generation performance comparison on different tasks

Figure 1. **(a) Overview of UniScene.** Given BEV layouts, UniScene facilitates versatile data generation, including semantic occupancy, multi-view video, and LiDAR point clouds, through an occupancy-centric hierarchical modeling approach. **(b) Performance comparison on different generation tasks.** UniScene delivers substantial improvements over SOTA methods in video, LiDAR, and occupancy generation.

Abstract

Generating high-fidelity, controllable, and annotated training data is critical for autonomous driving. Existing methods typically generate a single data form directly from a coarse scene layout, which not only fails to output rich data forms required for diverse downstream tasks but also struggles to model the direct layout-to-data distribution. In this paper, we introduce UniScene, the first unified framework for generating three key data forms — semantic occupancy, video, and LiDAR — in driving scenes. UniScene employs a progressive generation process that decomposes the complex task of scene generation into two hierarchical steps: (a) first generating semantic occupancy from a customized scene layout as a meta scene representation rich in both semantic and geometric information, and then (b) conditioned on occupancy, generating video and LiDAR data, respectively, with two novel transfer strategies of Gaussian-based Joint Rendering and Prior-guided Sparse Modeling. This occupancy-centric approach reduces the generation burden, especially for intricate scenes, while providing detailed intermediate representations for the subsequent generation stages. Extensive experiments demonstrate that UniScene outperforms previous SOTAs in the occupancy, video, and LiDAR generation, which also indeed benefits downstream driving tasks. The Project is available at https://arlo0o.github.io/uniscene/.

1. Introduction

The generation of high-quality driving scenes is a promising approach for autonomous driving (AD), as it helps mitigate the high resource demands associated with real-world data collection and annotation [19, 25, 34, 41]. Recent advancements in generative models, particularly diffusion models [19, 25, 34, 41], have made it possible to generate realistic synthetic data [43, 51, 65], facilitating the training of downstream tasks. Existing methods [13, 51, 54, 57, 75] typically use layout conditions derived from coarse geometric labels (*e.g.*, BEV maps and 3D bounding boxes) as input to guide scene generation. The resulting synthetic data are then leveraged to improve downstream tasks such as BEV segmentation [27, 29, 60] and 3D object detection [6, 10, 16, 37, 53].

Nevertheless, as depicted in Tab. 1, existing driving scene generation models predominantly focus on generating data in a single format (*e.g.*, RGB video) [13, 51, 54, 58, 69], without fully exploring the potential of generating data across multiple formats. This limits their applicability for a wide range of downstream tasks that require diverse sensor data (*i.e.*, RGB video, LiDAR) to ensure sufficient training for real-world scenarios [1, 3, 28, 30, 50]. Furthermore, previous methods attempt to capture the real-world distribution with a single-step layout-to-data modeling process given only coarse input conditions (*e.g.*, BEV layouts or 3D boxes) [13, 54, 58]. This direct learning strategy hinders the model's ability to capture the complex distributions inherent

*Share contribution
†Corresponding author

Method	Multi-view	Video	LiDAR	Occupancy
BEVGen [44]	✓	✗	✗	✗
BEVControl [65]	✓	✗	✗	✗
DriveDreamer [69]	✗	✓	✗	✗
Vista [14]	✗	✓	✗	✗
MagicDrive [13]	✓	✓	✗	✗
Drive-WM [54]	✓	✓	✗	✗
LiDARGen [74]	✗	✗	✓	✗
LiDARDiffusion [39]	✗	✗	✓	✗
LiDARDM [75]	✗	✗	✓	✗
OccSora [47]	✗	✗	✗	✓
OccLLama [56]	✗	✗	✗	✓
OccWorld [70]	✗	✗	✗	✓
UniScene (Ours)	✓	✓	✓	✓

Table 1. **Comparison of generation forms** with existing works.

in real-world driving scenes (*e.g.*, realistic geometry and appearance), often resulting in suboptimal performance, as illustrated in Fig. 1 (b). To address this challenge, recent methods in data-driven realistic generation [22, 36, 45] have sought to model complex distributions using intermediate representations as inductive biases, enabling the generation of high-quality results through hierarchical steps.

Thus, exploring an optimal intermediate representation for complex 3D generation tasks in autonomous driving is crucial for achieving high-quality outputs. Semantic occupancy, widely used in autonomous driving perception tasks, has recently been recognized as a superior scene representation due to its rich semantic and geometric information [17, 46, 50]. Building on this, recent advancements in volumetric generation [25, 39, 74, 75] highlight the significant potential of semantic occupancy, not only for depicting driving environments with enhanced 3D structural details but also for enabling more accurate and diverse scene generation. Compared to traditional 2D representations, such as BEV maps [13, 43, 51, 54, 58, 65, 69], 3D occupancy offers a richer and more detailed scene representation. Given these advantages, we argue that semantic occupancy is an ideal intermediate representation for decomposing complex driving scene generation tasks. It captures both semantic and geometric information, facilitating the generation of diverse data formats (*e.g.*, RGB video and LiDAR) while enhancing the flexibility and accuracy of the generation process.

To this end, we propose UniScene, a unified occupancy-centric framework designed for the versatile generation of semantic occupancy, video, and LiDAR data. As illustrated in Fig. 1 (a), UniScene adopts a decomposed learning paradigm and is structured hierarchically: it first generates 3D semantic occupancy from BEV scene layouts, and subsequently leverages this representation to facilitate the generation of video and LiDAR data. Specifically, in contrast to previous unconditional occupancy generation approaches [23, 24, 31], we use customized BEV layout sequences as controllable inputs to generate semantic occupancy sequences with spatial-temporal consistency. Unlike single-step layout-to-data learning methods [13, 44, 54, 58, 69], our approach utilizes the generated occupancy as an intermediate representation to guide the subsequent generation. To bridge the representation gap and ensure high-fidelity generation of video and LiDAR data, we introduce two novel representation transfer strategies: 1). A geometry-semantics joint rendering strategy, utilizing Gaussian Splatting [20, 72], to facilitate conditional video generation with detailed multi-view semantic and depth maps; 2). A prior-guided sparse modeling scheme for LiDAR data generation, which efficiently generates LiDAR points using occupancy-based priors. The contributions of our framework can be summarized as follows:

- We introduce UniScene, the first unified framework for versatile data generation in driving scenes. It jointly generates high-quality data across three formats: semantic occupancy, multi-view video, and LiDAR point clouds.
- We propose a decomposed conditional generation paradigm that progressively models complex driving scenes, effectively reducing the difficulty of generation. Fine-grained semantic occupancy is first generated as an intermediate representation, which then facilitates the subsequent generation of video and LiDAR data.
- To bridge the domain gap between occupancy and other data formats, we introduce two novel representation transfer strategies: one based on Gaussian Splatting rendering and the other leveraging a sparse modeling scheme.
- Extensive experiments across various generation tasks demonstrate that UniScene outperforms state-of-the-art methods in video, LiDAR, and occupancy generation. Moreover, the data generated by UniScene leads to significant enhancements in downstream tasks, including occupancy prediction, 3D detection, and BEV segmentation.

2. Related Work

Semantic Occupancy Representation. Semantic occupancy has emerged as a prominent 3D scene representation in autonomous driving. Current research primarily focuses on Semantic Occupancy Prediction (SOP). MonoScene [9] introduces a 3D SOP method based on a monocular camera. TPVFormer [18] presents a Tri-Perspective View framework for SOP. VPD [25] utilizes a generative diffusion model for 3D SOP. In the realm of occupancy forecasting, OccWorld [70] predicts future occupancy states based on prior occupancy, while OccLlama [56] incorporates Large Language Models (LLMs) to assist in future occupancy prediction. However, research on occupancy generation, particularly for temporal 3D occupancy sequence generation, is still limited. Recently, OccSora [47] employs a Diffusion Transformer (DiT) to generate occupancy, but the quality of the generated results still lags behind the ground truth.

Generation Models in Autonomous Driving. High-quality data is essential for training models in autonomous driving, which has led to a growing interest in driving scene

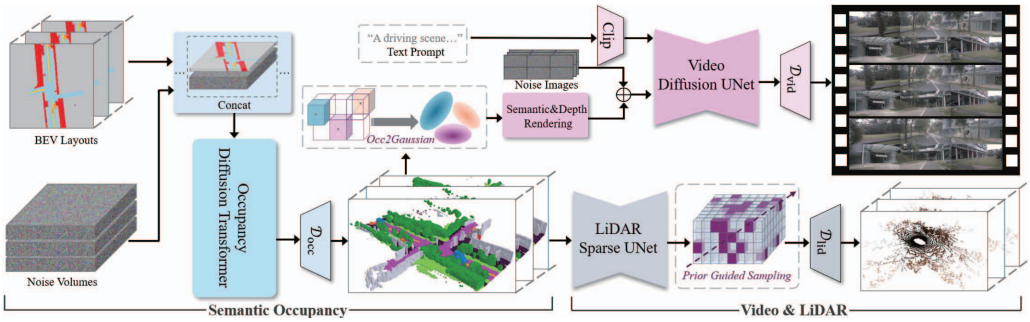


Figure 2. **Overall framework of the proposed method.** The joint generation process is organized into an occupancy-centric hierarchy: **I. Controllable Occupancy Generation.** The BEV layouts are concatenated with the noise volumes before being fed into the Occupancy Diffusion Transformer, and decoded with the Occupancy VAE Decoder $\mathcal{D}_{\text{occ}}$. **II. Occupancy-based Video and LiDAR Generation.** The occupancy is converted into 3D Gaussians and rendered into semantic and depth maps, which are processed with additional encoders as in ControlNet. The output is obtained from the Video VAE Decoder $\mathcal{D}_{\text{vid}}$. For LiDAR generation, the occupancy is processed via a sparse UNet and sampled with the geometric prior guidance, which is sent to the LiDAR head $\mathcal{D}_{\text{lid}}$ for generation.

generation tasks. One line of research employs Neural Radiance Fields (NeRF) and Gaussian Splatting (GS) techniques [61, 62, 64, 67] to synthesize novel perspectives, though these methods often suffer from limited scene diversity. With the rise of diffusion models, increasing attention has been given to generating driving images or videos, as seen in works like BEVGen [44], DriveDreamer [51], MagicDrive [13], and Panacea [57]. Some approaches also integrate world-model concepts into the generation process, such as Drive-WM [54], WoVoGen [33], and Vista [14]. In addition to generating images or videos, recent studies have explored the generation of LiDAR point clouds, including LidarDiffusion [39] and LidarDM [75]. Nevertheless, these methods primarily focus on single-form generation, overlooking the complementary nature of multi-modal data. In contrast, our versatile framework generates multiple forms of high-quality data, including occupancy, video, and LiDAR.

3. Methodology

In this section, we present UniScene, a unified framework designed to jointly generate three data forms: semantic occupancy, multi-view video, and LiDAR point cloud.

Overview. As shown in Fig. 2, we decompose the complex task of driving scene generation into an occupancy-centric hierarchy. Specifically, given multi-frame BEV layouts as conditions, UniScene first generates the corresponding semantic occupancy sequence with an Occupancy Diffusion Transformer (Sec. 3.1). The generated occupancy then serves as conditional guidance for subsequent video and LiDAR generation. For video generation, the occupancy is converted into 3D Gaussian primitives, which are then rendered into 2D semantic and depth maps to guide the video diffusion UNet (Sec. 3.2). For LiDAR generation, we propose a sparse modeling approach that combines a LiDAR sparse UNet with a ray-based sparse sampling strategy, guided by occupancy priors, to effectively produce LiDAR points (Sec. 3.3).

3.1. Controllable Semantic Occupancy Generation

Generating controllable and temporally consistent semantic occupancy is crucial in UniScene, as subsequent video and LiDAR generation depend on it. To address this, we introduce the Occupancy Diffusion Transformer (DiT), which takes BEV layout sequences as input, allowing users to easily edit and generate the corresponding occupancy sequences.

Temporal-aware Occupancy VAE. The occupancy VAE is designed to compress occupancy data into a latent space for computational efficiency. Unlike existing VQVAE-based approaches that rely on discrete tokenizers [56, 70], our method employs a VAE to encode occupancy sequences into a continuous latent space. It facilitates better preservation of spatial details, particularly under high compression ratios. Experimental evaluations are provided in Sec.4.1 and Tab.2.

Specifically, temporal information is only considered during the VAE decoding process, enabling greater flexibility as in [5]. In the VAE encoder, we transform the 3D occupancy $\mathbf{O} \in \mathbb{R}^{H\times W\times D}$ into a BEV representation $\hat{\mathbf{O}} \in \mathbb{R}^{H\times W\times DC'}$ following [70], where C' represents the dimension of the learnable class embedding. Subsequently, 2D convolutional layers and 2D axial attention layers are used to obtain a down-sampled continuous latent feature. The VAE decoder reconstructs the temporal latent feature $\boldsymbol{z}_{\text{occ}}^{\text{seq}} \in \mathbb{R}^{T\times C\times h\times w}$ into an occupancy sequence $\mathbf{O}^{\text{seq}} \in \mathbb{R}^{T\times H\times W\times D}$. It is built using 3D convolutional layers and 3D axial attention layers to capture temporal features. Similar to [70], we use cross-entropy loss $\mathcal{L}_{\text{CE}}$, Lovasz-softmax loss $\mathcal{L}_{\text{LS}}$, and KL divergence loss $\mathcal{L}_{\text{KL}}$ to train the VAE. The total loss for the occupancy VAE is:

$$\mathcal{L}_{\text{occ}}^{\text{vae}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{LS}} + \lambda_2 \mathcal{L}_{\text{KL}}, \tag{1}$$

where λ_1 and λ_2 are the respective loss weights. Additional details about the training setup for the occupancy VAE can be found in the supplementary materials.

Latent Occupancy DiT. The Latent Occupancy DiT learns to generate occupancy latent sequence from noise volumes with the condition of the BEV layout $\mathbf{B}$. Specifically, the BEV layouts are first concatenated with the noise volumes, which are further patchified before being fed into the Occupancy DiT. This explicit alignment strategy helps the model efficiently learn spatial relationships, enabling more precise

control over the generated sequences. The Occupancy DiT aggregates spatial-temporal information through a series of stacked spatial and temporal transformer blocks [35]. The loss function of Occupancy DiT is defined following [38]:

$$\mathcal{L}_{\text{occ}}^{\text{dit}} = \mathbb{E}\left[\sum_{i=1}^{T} \left\| \boldsymbol{f}_{\text{dit}}\left(\boldsymbol{z}_{\text{occ}}^{i}, \mathbf{B}^{i}\right) - \boldsymbol{\epsilon}_{\text{n}}^{i} \right\|^{2}\right], \tag{2}$$

where $\boldsymbol{f}_{\text{dit}}\left(\boldsymbol{z}_{\text{occ}}^{i}, \mathbf{B}^{i}\right)$ represents the model output, while $\boldsymbol{z}_{\text{occ}}^{i}$ represents the input noisy latent of i^{th} frame. $\boldsymbol{\epsilon}_{\text{n}}^{i}$ is the random noise sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. By incorporating temporal information, our occupancy diffusion model enables the generation of long-term consistent occupancy sequences (see Fig. 5 (a)). More architectural details about the Occupancy VAE and DiT can be found in the supplementary.

3.2. Video: Occupancy as Conditional Guidance

The video generation model is initialized with the pre-trained latent generation model of Stable Video Diffusion (SVD) [4], which is composed of a 3D video VAE and a video diffusion UNet. As depicted in Fig. 2, the video diffusion UNet takes occupancy-based rendering maps and text prompts as conditions to generate multi-view driving videos.

Gaussian-based Joint Rendering. The input semantic occupancy grids are jointly rendered into multi-view semantic and depth maps with forward Gaussian Splatting [20, 72]. The rendered maps bridge the representation gaps between the occupancy grids and the multi-view video, providing fine-grained semantic and geometric guidance to facilitate high-quality and consistent video generation. Rather than relying on the resource-intensive spatial-temporal attention mechanisms used in previous works [13, 59] for cross-view consistency, we retain the vanilla cross-attention from SVD [4] and ensure cross-view consistency through occupancy-based multi-view conditional guidance. Additional experimental evaluations are provided in Fig.9 and Tab.9.

Specifically, given input semantic occupancy with the shape of $\mathbb{R}^{H\times W\times D}$, we first convert it into a set of 3D Gaussian primitives $\mathcal{G} = \{G_i\}_{i=1}^{N}$ based on the center and semantic label of each occupancy grid. In this way, each Gaussian primitive contains attributes of position μ, semantic label s, opacity status α, and the covariance Σ calculated with the default rotation and voxel size scaling. Then, the depth map $\mathbf{D}$ and the semantic map $\mathbf{S}$ are rendered via tile-based rasterization [20] similar to color rendering:

$$\mathbf{D} = \sum_{i\in N} d_i \alpha_i' \prod_{j=1}^{i-1}\left(1-\alpha_j'\right), \tag{3}$$

$$\mathbf{S} = \texttt{argmax}\Big(\sum_{i\in N} \texttt{onehot}(s_i)\alpha_i' \prod_{j=1}^{i-1}\left(1-\alpha_j'\right)\Big), \tag{4}$$

where d_i denotes the depth value and α' is determined by the projected 2D Gaussian and the 3D opacity α as [20].

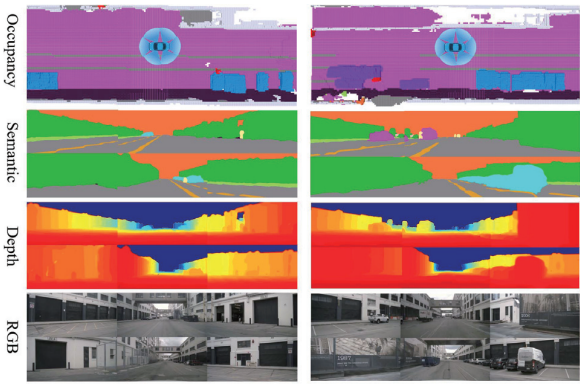


Figure 3. **Visualization** of the Gaussian-based joint rendering.

The visualization results of the rendered semantic and depth maps are shown in Fig. 3. Note that the road lines from the BEV layouts are projected onto the semantic occupancy, integrating the corresponding semantic information. These maps are fed into an encoder branch with residual connections and zero convolutions, similar to ControlNet [68], to leverage the pre-trained capabilities of the video diffusion UNet while preserving its inherent generative power.

Geometric-aware Noise Prior. To further enhance video generation quality, we introduce a geometry-aware noise prior strategy during the sampling process. It injects a dense appearance prior, similar to previous works [12, 52], while also incorporating explicit geometric awareness through the rendered depth map $\mathbf{D}$ to model regional correlations.

Specifically, the training noise in vanilla noise prior is:

$$\boldsymbol{\epsilon}_{\text{vid}}^{i} = \lambda \boldsymbol{z}_c + \boldsymbol{\epsilon}_{\text{n}}^{i}, \tag{5}$$

where $\boldsymbol{\epsilon}_{\text{vid}}^{i}$ represents the noise input of i^{th} video frame. $\boldsymbol{z}_c$ denotes the latent feature corresponding to the conditional video frame, obtained by encoding the conditional video frame using the 3D video VAE encoder. $\boldsymbol{\epsilon}_{\text{n}}^{i}$ is random noise sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. λ is the balancing coefficient.

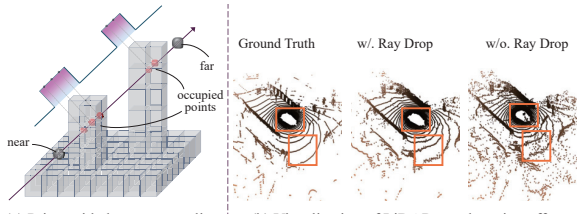
Nevertheless, in real-world scenarios, many regions within dynamic videos exhibit significant variations across multiple frames. The simple strategy described earlier does not account for correspondence modeling in these highly dynamic regions. To address this, we leverage the rendered depth map $\mathbf{D}$ to warp the appearance prior from the reference image to other images using depth-based reprojection [8, 55], enabling explicit geometric awareness. With this approach, we revisit Eq. 5 and optimize it as follows:

$$\boldsymbol{\epsilon}_{\text{vid}}^{i} = \lambda(\texttt{Warp}(\boldsymbol{z}_c, \mathbf{D}^{i}, \mathbf{K}, [\mathbf{R}_{0,i} \mid \mathbf{t}_{0,i}])) + \boldsymbol{\epsilon}_{\text{n}}^{i}, \tag{6}$$

where the depth-based `Warp` process [8, 55] projects the pixel $p_i = (u, v, 1)^T$ in the i^{th} latent feature z_i to its counterpart p_0 in the conditional latent feature z_c as:

$$p_0 = \mathbf{K} \cdot (\mathbf{R}_{0,i} \cdot (\mathbf{K}^{-1} \cdot p_i \cdot \mathbf{D}^{i}(u, v)) + \mathbf{t}_{0,i}). \tag{7}$$

where $\mathbf{K}$ is the camera intrinsics of the latent features, scaled from the original sensor parameters. $[\mathbf{R}_{0,i} \mid \mathbf{t}_{0,i}]$ is the transform matrix from target latent z_i to conditional latent z_c.



(a) Prior guided sparse sampling (b) Visualization of LiDAR ray-dropping effect

Figure 4. (a) **Sparse sampling** with occupancy-based prior guidance. (b) **Visualization** of the effect on LiDAR ray-dropping head.

$\mathbf{D}^i$ is the rendered depth map of i^{th} video frame. More details of the evaluation results can be found in Tab. 9 and the supplementary.

Video Training Loss. We define the video training loss similar to previous works [4, 14], which is formulated as:

$$\mathcal{L}_{\text{vid}} = \mathbb{E}\left[\sum_{i=1}^{T}\left(1-m^i\right)\cdot\left\|\boldsymbol{f}_{\text{vid}}\left(\boldsymbol{z}_{\text{vid}}^i, t, \boldsymbol{z}_c, \mathbf{D}^i, \mathbf{S}^i\right) - \boldsymbol{z}_0^i\right\|^2\right], \tag{8}$$

where $\boldsymbol{f}_{\text{vid}}\left(\boldsymbol{z}_{\text{vid}}^i, t, \boldsymbol{z}_c, \mathbf{D}^i, \mathbf{S}^i\right)$ represents the output of the video generation model. $\mathbf{D}^i$ and $\mathbf{S}^i$ denote the rendered depth map and semantic map of i^{th} video frame, respectively. t represents the input text prompt. $\boldsymbol{z}_0^i, \boldsymbol{z}_{\text{vid}}^i$ are the ground truth and input noisy latent, respectively. $\boldsymbol{z}_c$ is the conditional reference frame leveraged following SVD [4]. m is the one-hot mask used to select the condition frames. Note that we randomly select $\boldsymbol{z}_c$ to diminish the model's reliance on any specific conditioning frame.

3.3. LiDAR: Occupancy-based Sparse Modeling

As illustrated in Fig. 2, for LiDAR generation, the input occupancy is first encoded into sparse voxel features with a Sparse UNet [42], which are then used to produce LiDAR points via sparse sampling guided by occupancy priors.

Prior Guided Sparse Modeling. Given the inherent sparsity and detailed geometry of semantic occupancy, we propose a prior-guided sparse modeling approach to enhance computational efficiency (see the last column of Tab.5), by avoiding unnecessary computations on unoccupied voxels. The input semantic occupancy grids are first processed using a Sparse UNet[42] to aggregate contextual features. Next, we perform uniform sampling along the LiDAR rays, denoted as $\mathbf{r}$, to generate a sequence of points, represented as s.

As shown in Fig. 4 (a), to facilitate prior-guided sparse sampling, we assign a probability of 1 to points within occupied voxels and 0 to all other points, thus defining a probability distribution function (PDF). Subsequently, we resample n points $\{\mathbf{r}_i = o + s_i v \ (i = 1, ..., n)\}$ based on the PDF. Here o is the ray origin and v is the normalized ray direction.

LiDAR Head and Training Loss. Following the ray-based volume rendering techniques from previous works [2, 48, 66], the features of each resampled point are processed through a multi-layer perceptron (MLP) to predict the signed distance function (SDF) $f(s)$ and to compute the associated weights $w(s)$. These predictions and weights are then used to estimate the depth of the ray via volume rendering:

$$\beta_i = \max\left(\frac{\Phi_s(f(\mathbf{r}(s_i))) - \Phi_s(f(\mathbf{r}(s_{i+1})))}{\Phi_s(f(\mathbf{r}(s_i)))}, 0\right), \tag{9}$$

$$w(s_i) = \prod_{j=1}^{i-1}(1-\beta_j)\beta_i, \quad h = \sum_{i=1}^{n} w(s_i)s_i, \tag{10}$$

where $\Phi_s(x) = (1+e^{-sx})^{-1}$, h is the rendered depth value. To more accurately simulate the realistic LiDAR imaging process, we incorporate an additional reflection intensity head and a ray-dropping head. The reflection intensity head predicts the intensity of the object's reflection of the LiDAR laser beam along each ray. This prediction involves a weighted sum of the point features along the ray according to $w(s)$, followed by an MLP for estimation. The ray-dropping model estimates the probability of a ray not being captured by the LiDAR due to the failure in detecting reflected light, which is implemented with the same structure as the reflection intensity head. As shown in Fig. 4 (b), the ray-dropping head effectively eliminates noise points in the predictions. The training loss for LiDAR generation is composed of depth loss $\mathcal{L}_{\text{depth}}$, intensity loss $\mathcal{L}_{\text{inten}}$ and ray-dropping loss L_{drop}:

$$\mathcal{L}_{\text{lid}} = \mathcal{L}_{\text{depth}} + \lambda_1\mathcal{L}_{\text{inten}} + \lambda_2\mathcal{L}_{\text{drop}}, \tag{11}$$

where λ_1, λ_2 are balancing coefficients. More details including the training setup are provided in the supplementary.

4. Experiments

Our experiments are conducted on the challenging NuScenes benchmark [7]. We interpolate the semantic occupancy labels from the 2Hz key-frame annotations of the NuScenes-Occupancy dataset [50] to a higher frame rate of 12Hz, similar to previous studies [13, 49, 69]. For additional details, including dataset specifics, evaluation metrics, comparison baselines, model configurations, and further visualizations, please refer to the supplementary materials.

4.1. Main Results

Versatile Generation Ability. As shown in Fig. 5, the proposed UniScene facilitates versatile and controllable generation. For large-scale coherent generation, UniScene first produces long-term semantic occupancy from the given BEV layouts, which subsequently guides the generation of LiDAR point clouds and multi-view videos (Fig. 5(a)). Regarding controllable geometry editing, users can easily edit the BEV layouts, such as by removing cars (Fig. 5(b)), and the generation results can change accordingly. Furthermore, as depicted in Fig. 5(c), for attribute customization, UniScene allows for varied weather and lighting conditions in the generated videos through user-specified text prompts (*e.g.*, specifying weather or time of day).

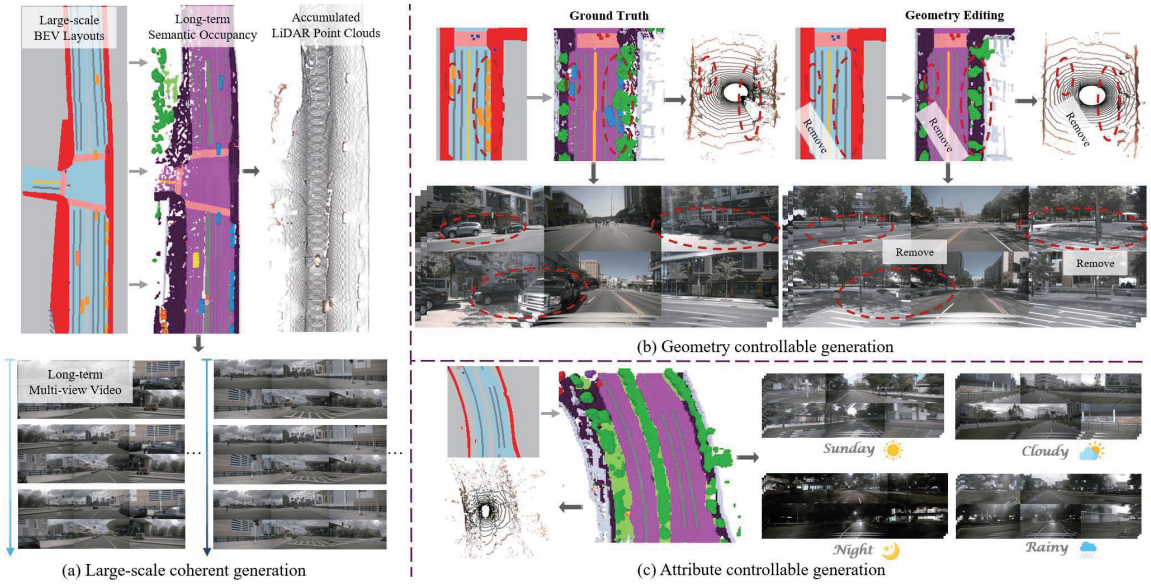


Figure 5. **Versatile generation ability of UniScene.** (a) Large-scale coherent generation of semantic occupancy, LiDAR point clouds, and multi-view videos. (b) Controllable generation of geometry-edited occupancy, video, and LiDAR by simply editing the input BEV layouts to convey user commands. (c) Controllable generation of attribute-diverse videos by changing the input text prompts.

Method	Compression Ratio ↑	mIoU ↑	IoU ↑
OccLLama (VQVAE) [56]	8	75.2	63.8
OccWorld (VQVAE) [70]	16	65.7	62.2
OccSora (VQVAE) [47]	**512**	27.4	37.0
Ours (VAE)	32	**92.1**	**87.0**
Ours (VAE)	**512**	72.9	64.1

Table 2. **Occupancy Reconstruction.** The compression ratio is calculated following the methodology outlined in OccWorld [70]. The top two performers are highlighted in **bold** and underlined.

Method	CFG	mIoU ↑	F3D ↓	MMD ↓
Ours-Gen.	4	20.51	205.78	11.60
	1	19.44	158.55	10.60
OccWorld [70]	-	17.13	145.65	9.89
Ours-Fore.	-	**31.76**	**43.13**	**2.86**

Table 3. **Occupancy Generation and Forecasting.** 'Ours-Gen.' and 'Ours-Fore.' denote our Generation model and Forecasting model, respectively. 'CFG' refers to the Classifier-Free Guidance.

Occupancy Reconstruction, Generation and Forecasting. The evaluation of occupancy is on the NuScenes-Occupancy key-frame (2Hz) validation set. Since precisely reconstructing the occupancy is vital for occupancy generation, we first compare our Occupancy VAE with the existing methods on reconstruction accuracy in Tab. 2. Compared to the discrete compression with VQVAE in previous works [47, 56, 70], our continuous compression with VAE achieves remarkable reconstruction performance even under the high compression ratio of 512, surpassing OccWorld [70] by 10.96% in mIoU. Note that the compression ratio of 512 is employed as the default setting in our model to ensure efficiency.

To further evaluate the effectiveness of our occupancy generation model, we implement a forecasting variant of our model as previous forecasting works [70] for comparison. Please refer to the supplementary for forecasting adaption details. The quantitative evaluation for occupancy generation ('Ours-Gen.') and forecasting ('Ours-Fore.') is shown in Tab. 3. Even with fewer reference occupancy frames ('Ours-Fore.': 2, OccWorld: 5), 'Ours-Fore' surpasses OccWorld by 70.39% in F3D and 71.08% in MMD, respectively. The qualitative results of occupancy forecasting are shown in Fig. 6. Our method can handle dynamic objects and sharp steering maneuvers compellingly in the generating process.

Method	Multi-view	Video	FID ↓	FVD ↓
BEVGen [44]	✓	✗	25.54	-
BEVControl [65]	✓	✗	24.85	-
DriveGAN [21]	✗	✓	73.40	502.30
DriveDreamer [69]	✗	✓	52.60	452.00
Vista [14]	✗	✓	6.90	89.40
WoVoGen [33]	✓	✓	27.60	417.70
Panacea [57]	✓	✓	16.96	139.00
MagicDrive [13]	✓	✓	16.20	-
Drive-WM [54]	✓	✓	15.80	122.70
Vista * [14]	✓	✓	13.97	112.65
Ours (Gen Occ)	✓	✓	6.45	71.94
Ours (GT Occ)	✓	✓	**6.12**	**70.52**

Table 4. **Video Generation.** We implement the multi-view variant of Vista* [14] with spatial-temporal attention [13, 59].

Method	MMD (10^{-4})↓	JSD ↓	Time (s)↓
LiDARDM [75]	3.51	0.118	45.12
Open3D [73]	8.15	0.149	2.39
Ours (Gen Occ)	2.40	0.108	0.47
Ours (GT Occ)	**1.53**	**0.072**	**0.25**

Table 5. **LiDAR Generation.** We include the semantic occupancy generation time for a fair comparison.

Video Generation Results. The quantitative comparison of

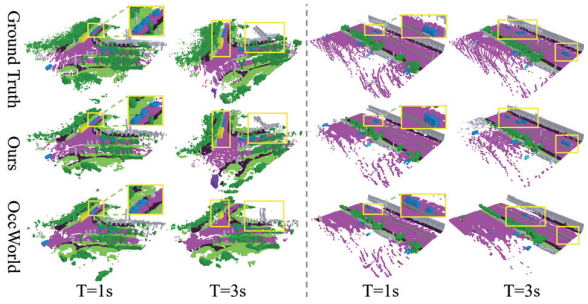


Figure 6. **Qualitative evaluation** for occupancy forecasting. Our method can compellingly handle sharp steering maneuvers and dynamic objects with temporal consistency.

video generation is illustrated in Tab. 4. The 'Gen Occ' and 'GT Occ' represent the occupancy conditions generated from our framework and ground truth, respectively. Our method supports multi-view video generation and outperforms all the other methods, achieving 71.94 FVD with generated occupancy and 70.52 FVD with ground truth occupancy, respectively. Note that the competitive method of Vista [14] only supports single-view video generation, here we implement the multi-view variant with spatial-temporal attention following [13, 59] for a fair comparison.

As shown in Fig. 7, we compare our video generation results with the video generation model of MagicDrive [13]. Our approach demonstrates an obvious improvement in video generation quality, particularly in the structure quality of objects and cross-view consistency. The notable enhancement is attributed to the conditional guidance derived from occupancy-based semantic and geometric rendering maps.

LiDAR Generation Results. We compare our LiDAR generation model with Open3D [73] and LiDARDM [75] on the NuScenes validation set. For Open3D, we utilize the library's hard ray-casting function to convert the ground truth occupancy grids into corresponding LiDAR point clouds. LiDARDM is implemented using the official repository and trained with the same setting as our model on NuScenes. The 'Gen Occ' and 'GT Occ' are the same as the video generation evaluation. As shown in Tab. 5, our method achieves the best generation performance, surpassing LiDARDM by 31.62% in MMD. Moreover, our approach demonstrates significant advantages in inference speed over other methods, which stem from our efficient sparse modeling scheme. The qualitative results are shown in Fig. 8. Compared to LiDARM and Open3D, our method shows significant superiority in generating precious scene layouts and clear object details.

Downstream Task Evaluation. UniScene can produce multi-modal augmented data with corresponding annotations, thereby enhancing training for downstream perception tasks. We augment an equal quantity of images and LiDAR point clouds as in the original NuScenes dataset, ensuring consistent training iterations and batch sizes for fair comparisons as previous works [13, 44]. Unlike existing works [13, 44] that only produce augmented images,

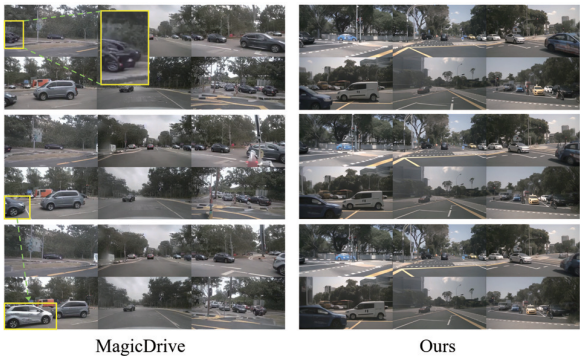


Figure 7. **Qualitative evaluation** for video generation. Our method excels in object structure quality and cross-view consistency.

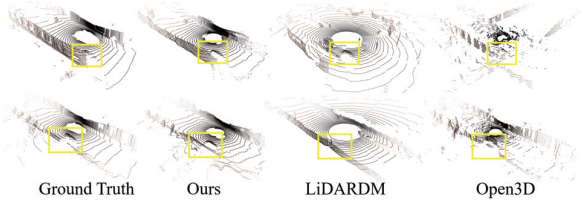


Figure 8. **Qualitative evaluation** for LiDAR generation. Our method generates precious scene layouts and clear object details.

we generate different data forms to enable the training of comprehensive downstream multi-modal perception models. For the Occupancy Prediction task in Tab. 6, UniScene significantly improves the baseline of CONet [50] in different modality settings (*e.g.*, Camera (C), LiDAR (L), Camera+LiDAR (C&L)), which stems from the high fidelity generation with precious occupancy-based conditional guidance of fine-grained geometry and semantics. For the BEV Segmentation and Object Detection tasks in Tab. 7, we employ CVT [71]) and BEVFusion [32] as baselines following [13]. Our method significantly surpasses other methods for data augmentation, highlighting the superior generation quality.

4.2. Ablation Studies

Effect of Designs in Occupancy Generation Model. We conduct extensive ablation studies to validate the effectiveness of the occupancy generation model components, as shown in Tab. 8. Incorporating temporal information in the occupancy VAE decoder with 3D axial attention boosts occupancy sequence generation fidelity, particularly reducing the MMD metric by 38.01%. The temporal attention layer in Occupancy DiT effectively improves the generation results, boosting the F3D metric by 10.29%. The spatial attention layer in Occupancy DiT brings significant enhancement in generation fidelity, reducing the F3D metric by 39.26%.

Effect of Designs in Video Generation Model. As shown in Tab. 9 and Fig. 9, we conduct extensive ablation studies to validate the effectiveness of the video generation model components. For the setting of 'w/. Spatial-temporal Attention',

Method	Input	IoU ↑	mIoU ↑
MonoScene [9]	C	18.4	6.9
TPVFormer [18]	C	15.3	7.8
Baseline-C [50]	C	20.1	12.8
w/. Vista* [14]	C	20.8+0.7	13.1+0.3
w/. MagicDrive [13]	C	21.8+1.7	13.9+1.1
w/. Ours-C	C	**28.6**+8.5	**16.5**+3.7
LMSCNet [40]	L	27.3	11.5
JS3C-Net [63]	L	30.2	12.5
Baseline-L [50]	L	30.9	15.8
w/. Ours-L	L	**33.1**+2.2	**19.3**+3.5
3DSketch [11]	$C\&L^D$	25.6	10.7
AICNet [26]	$C\&L^D$	23.8	10.6
Baseline-M [50]	$C\&L$	29.5	20.1
w/. Ours-M	$C\&L$	**35.4**+5.9	**23.9**+3.8

Table 6. **Comparison about training support for semantic occupancy prediction** (Baseline as CONet [50]). The "C", "L", and "L^D" denote the camera, LiDAR, and depth projected from LiDAR.

Method	Input	BEV Segmentation		3D Object Detection	
		Road mIoU ↑	Vehicle mIoU ↑	mAP ↑	NDS ↑
Baseline [32, 71]	C	74.30	36.00	32.88	37.81
w/. BEVGen [44]	C	71.90-2.40	36.60+0.60	-	-
w/. Vista* [14]	C	76.62+2.32	37.71+1.71	34.04+1.16	38.60+0.79
w/. MagicDrive [13]	C	79.56+5.26	40.34+4.34	35.40+2.52	39.76 +1.95
w/. Ours	C	**81.69**+7.39	**41.62**+5.62	**36.50**+3.62	**41.22**+3.41
Baseline-M [32]	$C\&L$	-	-	65.40	69.59
w/. Ours-M	$C\&L$	-	-	**68.53**+3.13	**72.20**+2.61

Table 7. **Comparison about training support for BEV segmentation** (Baseline as CVT [71]) and **3D object detection** (Baseline as BEVFusion [32]).

we remove the occupancy-based rendering maps and employ spatial-temporal attention as [13, 59]. Our results indicate that the occupancy-based semantic and geometric rendering maps are more effective in enhancing video generation quality. Moreover, as illustrated in Fig. 9, our strategy markedly improves the cross-view consistency with high-fidelity structures. The rendered semantic and depth maps from the semantic occupancy benefit the fidelity of image and video generation significantly with detailed prior guidance, boosting 34.66% and 31.15% in terms of FVD, respectively. The geometric-aware noise prior improves the video generation performance obviously, reducing FVD from 87.52 to 70.52.

Effect of Designs in LiDAR Generation Model. The ablation studies on the proposed LiDAR generation model are illustrated in Tab. 10. For the setting of 'w/o. Sparse UNet', we replace the Sparse UNet with a single layer of submanifold convolution [15]. As we can see, the Sparse UNet efficiently improves the generation performance (-25.77% in JSD) with a relatively small memory increment (+1.63%). To implement the setting of 'w/o. Sparse Sampling', we replace it with uniform sampling following NeuS [48]. Our sparse sampling strategy brings a significant computational consumption decrease (-58.94%) with better performance (-4.00% in JSD). Moreover, the ray-dropping head reduces MMD and JSD by 25.92% and 25.00%, respectively. We attribute such obvious improvement to the capability of realistic LiDAR ray-dropping phenomenon simulation.

Method	mIoU ↑	F3D ↓	MMD ↓
Ours	**19.44**	**158.55**	**10.60**
w/o. VAE 3D Axial Attention	18.77	167.91	17.10
w/o. DiT Temporal Attention	17.63	176.74	11.35
w/o. DiT Spatial Attention	10.29	261.03	18.59

Table 8. **Ablation** for designs in the occupancy generation model.

Method	FID↓	FVD↓
Ours	**6.12**	**70.52**
w/. Spatial-temporal Attention	12.72	110.87
w/o. Rendered Semantic Map	11.72	107.92
w/o. Rendered Depth Map	10.17	102.42
w/o. Depth-based Noise Prior	7.23	87.52

Table 9. **Ablation** for designs in the video generation model.

Method	MMD (10^{-4}) ↓	JSD ↓	Time (s)↓	Memory (GB)↓
Ours	**1.53**	**0.072**	0.25	6.84
w/o. Sparse UNet	2.88	0.097	**0.21**	6.73
w/o. Sparse Sampling	1.69	0.075	0.30	16.66
w/o. Ray-dropping Head	3.25	0.100	0.25	**5.05**

Table 10. **Ablation** for designs in the LiDAR generation model.

w/. spatial-temporal attention w/. occupancy-based conditional guidance

Figure 9. **Visualization** of the effect of occupancy-based conditional guidance. Our approach significantly enhances cross-view consistency, resulting in high-fidelity structures, while spatial-temporal attention fails to achieve similar results.

5. Conclusion

In this paper, we introduce UniScene, a unified framework designed to generate high-fidelity, controllable, and annotated data for autonomous driving applications. By decomposing the complex scene generation task into two hierarchical steps, UniScene progressively produces semantic occupancy, video, and LiDAR data. Extensive experiments show that UniScene surpasses current SOTAs in all three data types and enhances extensive downstream tasks.

Acknowledgments

This work was supported by NSFC 62302246, ZJNSFC under Grant LQ23F010008, Ningbo Grant 2023Z237 & 2024Z284 & 2024Z289 & 2023CX050011, and supported by High Performance Computing Center at Eastern Institute of Technology and Ningbo Institute of Digital Twin.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, 2022. 1

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023. 5

[3] Julie Stephany Berrio, Mao Shan, Stewart Worrall, and Eduardo Nebot. Camera-lidar integration: Probabilistic sensor fusion for semantic mapping. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):7637–7652, 2021. 1

[4] 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. 4, 5

[5] 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 *CVPR*, 2023. 3

[6] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018. 1

[7] 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 multimodal dataset for autonomous driving. In *CVPR*, 2020. 5

[8] Changjiang Cai, Pan Ji, Qingan Yan, and Yi Xu. Riav-mvs: Recurrent-indexing an asymmetric volume for multi-view stereo. In *CVPR*, 2023. 4

[9] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022. 2, 8

[10] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv preprint arXiv:2306.04607*, 2023. 1

[11] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *CVPR*, 2020. 8

[12] Huan-ang Gao, Mingju Gao, Jiaju Li, Wenyi Li, Rong Zhi, Hao Tang, and Hao Zhao. Scp-diff: Photo-realistic semantic image synthesis with spatial-categorical joint prior. *arXiv preprint arXiv:2403.09638*, 2024. 4

[13] 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. In *ICLR*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[14] 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, 7, 8

[15] Benjamin Graham and Laurens Van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 8

[16] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 1

[17] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, 2023. 2

[18] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, 2023. 2, 8

[19] Hao Jiang and Yadong Mu. Conditional diffusion process for inverse halftoning. *NeurIPS*, 35:5498–5509, 2022. 1

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 2, 4

[21] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*, 2021. 6

[22] Animesh Koratana, Daniel Kang, Peter Bailis, and Matei Zaharia. Lit: Learned intermediate representation training for model compression. In *ICML*. PMLR. 2

[23] Jumin Lee, Woobin Im, Sebin Lee, and Sung-Eui Yoon. Diffusion probabilistic models for scene-scale 3d categorical data. *Workshop on Image Processing and Image Understanding*, 2023. 2

[24] Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, and Sung-Eui Yoon. Semcity: Semantic scene generation with triplane diffusion. In *CVPR*, 2024. 2

[25] Bohan Li, Yasheng Sun, Jingxin Dong, Zheng Zhu, Jinming Liu, Xin Jin, and Wenjun Zeng. One at a time: Progressive multi-step volumetric probability learning for reliable 3d scene perception. In *AAAI*, 2024. 1, 2

[26] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *CVPR*, 2020. 8

[27] Wenyi Li, Haoran Xu, Guiyu Zhang, Huan-ang Gao, Mingju Gao, Mengyu Wang, and Hao Zhao. Fairdiff: Fair segmentation with point-image diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 1

[28] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *CVPR*, 2022. 1

[29] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023. 1

[30] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *NeurIPS*, 2022. 1

[31] Yuheng Liu, Xinke Li, Xueting Li, Lu Qi, Chongshou Li, and Ming-Hsuan Yang. Pyramid diffusion for fine 3d large scene generation. *ECCV*, 2024. 2

[32] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*. IEEE, 2023. 7, 8

[33] Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. *arXiv preprint arXiv:2312.02934*, 2023. 3, 6

[34] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, pages 2837–2845, 2021. 1

[35] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 4

[36] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020. 2

[37] Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*, 2023. 1

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 4

[39] Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with lidar diffusion models. In *CVPR*, 2024. 2, 3

[40] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, 2020. 8

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1

[42] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2647–2664, 2020. 5

[43] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 2024. 1, 2

[44] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 2024. 2, 3, 6, 7, 8

[45] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *NeurIPS*, 2023. 2

[46] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *ICCV*, pages 8406–8415, 2023. 2

[47] Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*, 2024. 2, 6

[48] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 5, 8

[49] Xiaofeng Wang, Zheng Zhu, Yunpeng Zhang, Guan Huang, Yun Ye, Wenbo Xu, Ziwei Chen, and Xingang Wang. Are we ready for vision-centric driving streaming perception? the asap benchmark. *arXiv preprint arXiv:2212.08914*, 2022. 5

[50] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. *ICCV*, 2023. 1, 2, 5, 7, 8

[51] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *ECCV*, 2024. 1, 2, 3

[52] Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, et al. Microcinema: A divide-and-conquer approach for text-to-video generation. In *CVPR*, 2024. 4

[53] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, et al. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. *CVPR*, 2024. 1

[54] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *CVPR*, 2024. 1, 2, 3, 6

[55] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *CVPR*, 2021. 4

[56] Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024. 2, 3, 6

[57] 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, 2023. 1, 3, 6

[58] 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 *CVPR*, 2024. 1, 2

[59] 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 *ICCV*, 2023. 4, 6, 7, 8

[60] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. In *NeurIPS*, 2023. 1

[61] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, Yuxin Huang, Xiaoyu Ye, Zike Yan, Yongliang Shi, Yiyi Liao, and Hao Zhao. Mars: An instance-aware, modular and realistic simulator for autonomous driving. *CICAI*, 2023. 3

[62] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *ECCV*. 3

[63] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI*, 2021. 8

[64] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 3

[65] 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. 1, 2, 6

[66] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *CVPR*, 2023. 5

[67] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *AAAI*, 2024. 3

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 4

[69] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024. 1, 2, 5, 6

[70] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023. 2, 3, 6

[71] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, 2022. 7, 8

[72] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *CVPR*, 2024. 2, 4

[73] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 6, 7

[74] Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point cloud. In *ECCV*, 2022. 2

[75] Vlas Zyrianov, Henry Che, Zhijian Liu, and Shenlong Wang. Lidardm: Generative lidar simulation in a generated world. *arXiv preprint arXiv:2404.02903*, 2024. 1, 2, 3, 6, 7