OmniGen: Unified Multimodal Sensor Generation for Autonomous Driving

Tang Tao*
Shenzhen Campus of Sun Yat-sen University
Shenzhen, China
trent.tangtao@gmail.com

Enhui Ma
Westlake University
Hangzhou, China
maenhui@westlake.edu.cn

Xia Zhou
Li Auto Inc.
Beijing, China
zhouxia@lixiang.com

Letian Wang
The University of Toronto
Toronto, Canada
letianwang0@gmail.com

Tianyi Yan
Li Auto Inc.
Beijing, China
tianyi.yan123@gmail.com

Xueyang Zhang
Li Auto Inc.
Beijing, China
zhangxueyang@lixiang.com

Kun Zhan
Li Auto Inc.
Beijing, China
zhankun@lixiang.com

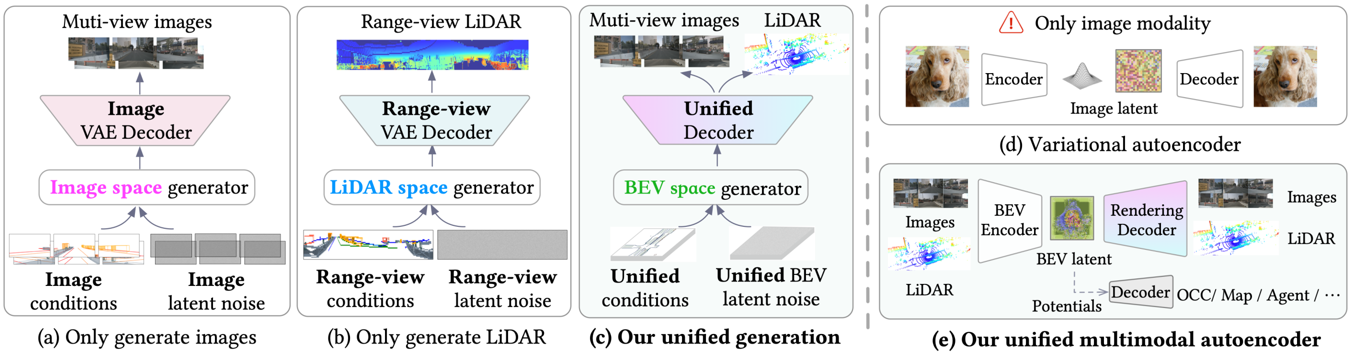
Peng Jia
Li Auto Inc.
Beijing, China
jiapeng@lixiang.com

Xianpeng Lang
Li Auto Inc.
Beijing, China
langxianpeng@lixiang.com

Jia-Wang Bian
Bytedance Seed
San Jose, United States
jiawang.bian@gmail.com

Kaicheng Yu
Westlake University
Hangzhou, China
kaicheng.yu.yt@gmail.com

Xiaodan Liang†
Shenzhen Campus of Sun Yat-sen University
Shenzhen, China
hxdliang328@gmail.com



(a) Only generate images

(b) Only generate LiDAR

(c) Our unified generation

(d) Variational autoencoder

(e) Our unified multimodal autoencoder

Figure 1: (a)(b) Existing approaches primarily focus on single-modality generation on its own space, leading to inefficiencies and misalignment in multimodal sensor data. (c) Our OmniGen, a unified multimodal sensor generation framework. (d) Traditional VAE, which only supports images. (e) Our UAE, a unified multimodal autoencoder, leverages the unified BEV latent.

Abstract

Autonomous driving has seen remarkable advancements, largely driven by extensive real-world data collection. However, acquiring diverse and corner-case data remains costly and inefficient. Generative models have emerged as a promising solution by synthesizing realistic sensor data. However, existing approaches primarily focus on single-modality generation, leading to inefficiencies and misalignment in multimodal sensor data. To address these challenges, we propose OmniGen, which generates aligned multimodal sensor data in a unified framework. Our approach leverages a shared Bird's Eye View (BEV) space to unify multimodal

*Work done during an internship at Li Auto Inc.
†Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754772

features and designs a novel generalizable multimodal reconstruction method, UAE, to jointly decode LiDAR and multi-view camera data. UAE achieves multimodal sensor decoding through volume rendering, enabling accurate and flexible reconstruction. Furthermore, we incorporate a Diffusion Transformer (DiT) with a ControlNet branch to enable controllable multimodal sensor generation. Our comprehensive experiments demonstrate that OmniGen achieves desired performances in unified multimodal sensor data generation with multimodal consistency and flexible sensor adjustments.

CCS Concepts

• **Computing methodologies → Computer vision**.

Keywords

OminiGen, Multimodal Sensor Generation, Autonomous Driving

ACM Reference Format:
Tang Tao, Enhui Ma, Xia Zhou, Letian Wang, Tianyi Yan, Xueyang Zhang, Kun Zhan, Peng Jia, Xianpeng Lang, Jia-Wang Bian, Kaicheng Yu, and Xiaodan Liang. 2025. OmniGen: Unified Multimodal Sensor Generation for Autonomous Driving. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754772

1 Introduction

Autonomous driving has made remarkable progress in recent years, driven by large-scale real-world data collected from diverse environments. However, the high cost and inefficiency of acquiring real-world data, especially in corner-case scenarios, limit the scalability of this process. Meanwhile, generative models [2, 37, 62] have gained significant attention for their ability to learn data distributions and synthesize realistic content, achieving remarkable success in image generation. Consequently, using generative models to synthesize desired sensors has become a de-facto standard in autonomous driving to address the data scarcity issue. For camera data generation, researchers directly fine-tune existing image generation models, such as ControlNet [62], with specific driving scene layouts or textual descriptions as conditions to generate scene images. Moreover, beyond cameras, LiDAR sensors play a crucial role in practical autonomous driving systems for accurate perception and planning, providing reliable 3D environmental measurements by capturing point clouds. For LiDAR data generation, to leverage existing image generation models, researchers first convert point clouds into range-view pseudo images and then adapt the whole image generation pipeline, *i.e.*, the autoencoder and diffusion model, to range-view space to generate pseudo LiDAR images, which are subsequently converted back into LiDAR points.

However, as illustrated in Fig. 1, existing driving scene generation models primarily focus on single-modality sensor data, while the unified generation of multimodal data remains unexplored. Unified multimodal sensor generation offers several advantages: **Improved efficiency** – generating both modalities simultaneously eliminates the need for separate pipelines (*e.g.*, the data processing, model training, and model updates). **Better sensor alignment** – Independently generating sensor data is difficult to align across different sensors, making it challenging for downstream multimodal models to utilize effectively. Nevertheless, achieving a unified multimodal generation poses significant challenges. Camera generation models operate in the image latent space, where conditions are projected into the perspective image view, while LiDAR generation models generate data in the range-view latent space with conditions projected accordingly. Fusing these distinct generation spaces into a unified representation, while ensuring it is controllable under a unified condition, is a non-trivial problem.

In this paper, we introduce OmniGen, a unified multimodal sensor generation framework for autonomous driving. We address the challenges of unified sensor generation by breaking the problem down into several key steps: **1) Establishing a unified representation space**: Inspired by prior multimodal perception research [24, 29], such as BEVFusion [25], we unify multimodal features in a shared Bird's Eye View (BEV) space, which provides a global scene context and aligns well with conditions such as textual descriptions or road sketches. **2) Decoding multimodal sensor data from the unified BEV space**: Drawing inspiration from generalizable NeRF approaches, *e.g.,* PixelNeRF [61], we leverage volume rendering to render sensor data. While previous generalizable NeRF methods primarily focus on object-level reconstruction and single-modality image rendering, their application to large-scale and multimodal autonomous driving scenes remains limited. Thanks to the unified BEV representation, in this work, we introduce UAE, the first generalizable LiDAR-camera multimodal reconstruction method for autonomous driving scenes. To decode multimodal data, UAE first upsamples the BEV features into 3D voxel features and then renders sensor features by sampling the voxel features. Subsequently, a carefully designed feature decoder is incorporated to map the rendered features to the corresponding sensor data, preserving and enhancing high-frequency details for improving reconstruction quality. Compared to traditional VAEs [20], UAE naturally offers multiple benefits from its generalizable reconstruction capability. For instance, autonomous driving scenes typically involve multiple surrounding images. Previous works process each view independently using VAEs and apply attention modules as soft constraints to enforce cross-view consistency. In contrast, UAE inherently supports multi-view consistent rendering by leveraging unified BEV features as a global scene constraint. Similarly, multimodal sensor data rendered from the shared BEV features also maintains consistency across different modalities. Moreover, UAE allows for the adjustment of sensor parameters (both intrinsic and extrinsic) during the decoding process. This enables effortless camera control, which previously required complex designs and training of generation models to achieve [27, 59]. **3) Generating latent BEV features for multimodal sensor data**: To enable generative models to produce multimodal sensor data, we enhance UAE with a Vector Quantization (VQ) module, and leverage it as our multimodal autoencoder for unified generation. For the generative model, we employ the ControlNet-Transformer architecture, which incorporates ControlNet into the powerful Diffusion Transformer (DiT) model. Moreover, we employ comprehensive scene conditions, i.e, scene textual descriptions, BEV road sketches, and 3D boxes, enabling more fine-grained and precise control of the generative model to generate the desired latent BEV features effectively. Overall, we achieve an end-to-end unified multimodal sensor generation framework. Given driving scene conditions, our

OmniGen can generate aligned multimodal sensor data. Moreover, we explore various multimodal architecture designs and provide valuable investigations into unified multimodal generation.

Through comprehensive experiments, we validate the effectiveness of our approach in generating multimodal sensor data within a unified framework. Specifically, our UAE module achieves state-of-the-art performance over previous generalizable reconstruction, and our OmniGen achieves comparable results with previous specialized single-modality methods and effectively generates multimodal data with cross-modality alignment and flexible sensor control, further enhancing downstream autonomous driving tasks.

Overall, our contributions are summarized as follows:

- We introduce OmniGen, a unified multimodal sensor generation framework that enables the controllable generation of aligned LiDAR and multi-view camera data.
- We propose UAE, a generalizable multimodal reconstruction method, which serves as an efficient multimodal autoencoder for encoding and decoding multimodal data in a unified space while allowing multimodal and multi-view consistency and flexible sensor adjustments.
- We demonstrate the effectiveness of our method quantitatively and qualitatively through extensive experiments conducted on multiple scenes.

2 Related Work

2.1 Camera Generation Models

Recent advancements in generative models, particularly in diffusion models [14, 62], have inspired the generation of high-fidelity and controllable driving scenes. Previous works [8, 11, 12, 19, 30, 31, 33, 48, 49, 51] have fine-tuned image generation models on driving data, incorporating various control signals such as maps, object bounding boxes, and textual descriptions to generate diverse driving scenarios. Specifically, BEVGen [42] first introduces the use of a BEV map as a condition for generating multi-view street images. BEVControl [57] further proposes a two-stage generation pipeline that integrates cross-view attention. MagicDrive [9] highlights 3D geometric information, encoding boxes and road maps separately to enable more fine-grained control. DriveDreamer [47] and DriveDreamer-2 [64] generate multi-view video data based on diverse control signals. Recently, researchers have been further refining driving image generation. For example, some works [21, 22, 54], *e.g.*, InfiniCube [32], adopt semantic occupancy as an intermediate representation to improve generation quality. Others [10, 32, 67] incorporate additional reconstruction modules to synthesize 4D driving scenes. Moreover, some studies [27, 59] integrate camera pose parameters into the generator to achieve more precise control. In this work, beyond scene images, we aim to jointly generate multimodal sensor data within a unified framework.

2.2 LiDAR Generation Models

Generative models also provide a promising alternative for creating realistic LiDAR point clouds without physics-based platforms. Early approaches, such as LiDARVAE and LiDARGAN [3], employ VAE or GANs for LiDAR cloud generation, but the realism achieved in their results is relatively limited. Following, UltraLiDAR [53] utilizes VQ-VAE [44] to generate voxelized LiDAR point clouds, while LidarDM [69] employs a map-conditioned diffusion model to generate scene meshes, which are then raycast to produce LiDAR scans. Meanwhile, with the advanced diffusion models, many works explore LiDAR generation based on range-view image representations. LiDARGen [68] firstly applies a diffusion model on range-view images, leveraging progress in image diffusion models. R2DM [35] designs a more mature diffusion framework, achieving significant performance improvements. RangeLDM [15] further optimizes efficiency and quality by compressing range-view data into a latent space before diffusion. More recently, LiDM [40] and Text2LiDAR [50] explore conditional LiDAR generation using conditions such as text, bounding boxes, and maps. Despite these advancements, state-of-the-art LiDAR generation methods operate in range-view space, which is challenging to unify with the image space used for camera sensor generation. In this work, we address this issue by unifying multimodal features within a shared BEV space, enabling unified multimodal sensor generation.

2.3 Generalizable NeRFs

Latent diffusion models fundamentally operate in the latent space of autoencoders. However, commonly used autoencoders, such as VAE [20] and VQ-VAE [44], do not inherently support multimodal sensor data. On the other hand, generalizable NeRFs [5, 26, 61, 63] replace the costly per-scene optimization with a single feedforward pass. These models take several images as input and generate corresponding image outputs, effectively functioning as a more versatile autoencoder. While previous generalizable NeRF methods have primarily focused on object-level reconstruction, only DistillNeRF [45] recently explored generalizable scene reconstruction for driving scenes. However, it relies on multiple complex modules, such as distillation from offline NeRFs, distillation from foundation models, hierarchical octree representation, and integration of depth features from DepthAnything [58], and it remains limited to image rendering. A unified and end-to-end multimodal generalizable NeRF—or a multimodal autoencoder—remains unexplored. In this work, we introduce UAE, a generalizable multimodal reconstruction method for driving scenes. Furthermore, we enhance UAE with a Vector Quantization (VQ) module and leverage it as a multimodal autoencoder for unified sensor data generation.

3 OmniGen

In this section, we introduce OmniGen in detail. We first give an overview in Section 3.1. Then, we introduce the UAE in Section 3.2 and unified LiDAR-Camera generation in Section 3.3.

3.1 Overview

As illustrated in Fig. 2, our OmniGen consists of a multimodal autoencoder and a multimodal generator. The multimodal autoencoder, UAE, first encodes the camera and LiDAR sensor data into a unified BEV space, then decodes the unified BEV features to multimodal sensor data by volume rendering. The multimodal generator adopts ControlNet-Transformer architecture, which incorporates ControlNet into the Diffusion Transformer (DiT) model, enabling fine-grained and precise control. Given textual descriptions, BEV road sketches, and 3D boxes as scene conditions input, the generator generates the desired latent BEV features effectively.

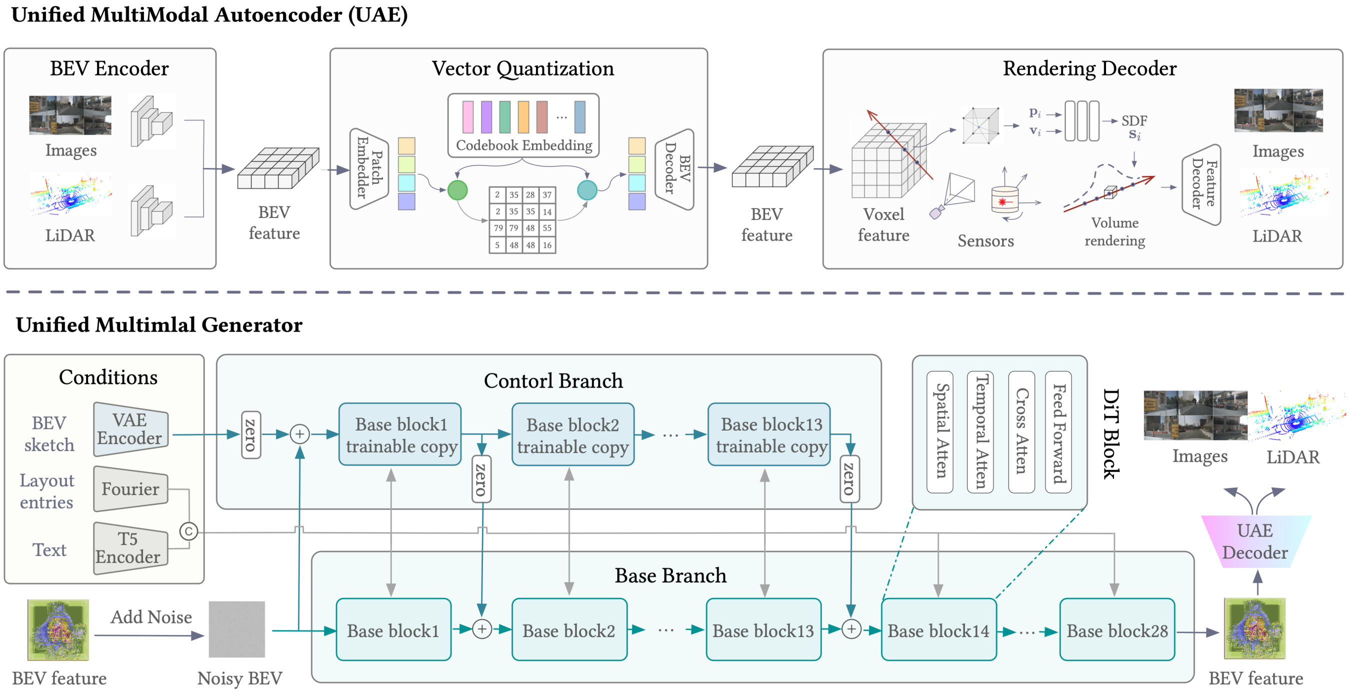


Figure 2: Our OmniGen framework consists of two main components: the unified multimodal autoencoder and the unified multimodal generator. The UAE is composed of three parts: a multimodal BEV encoder, a Vector Quantization module, and a multimodal rendering decoder. The unified multimodal generator includes two branches: the base branch and the control branch, which take multiple scene conditions as input to generate BEV latent representations.

3.2 Unified Multimodal Autoencoder

As commonly used autoencoders, *e.g.*, VAE [20] and VQ-VAE [44], do not support multimodal sensor data, we propose a novel unified multimodal autoencoder based on volume rendering and dub it as UAE. The UAE contains three parts: a multimodal BEV encoder, a multimodal rendering decoder, and a Vector Quantization module.

Multimodal BEV Encoder. The multimodal BEV encoder converts different modalities into a unified BEV space while preserving as much sensor-specific information as possible. Specifically, following prior works [23, 25], for multi-view images from the camera sensors, we adopt the Lift-Splat-Shoot (LSS) view transformation [38] to lift 2D features into the 3D volume features, denoted as $\mathbf{V}_C \in \mathbb{R}^{X \times Y \times Z \times C}$. For the LiDAR sensor, the point encoder firstly learns a parameterized voxelization [66] of the raw point clouds and then utilizes sparse 3D convolution networks [55] for efficient feature extraction. We follow UVTR [23] to directly retain the height dimension in the point encoder to obtain LiDAR 3D volume features, $\mathbf{V}_L \in \mathbb{R}^{X \times Y \times Z \times C}$. Then, $\mathbf{V}_C$ and $\mathbf{V}_L$ are summed and passed through a projection layer to enhance the fused voxel representation, forming the unified voxel space $\mathbf{V}_U \in \mathbb{R}^{X \times Y \times Z \times C}$. Finally, we adopt the Spatial-to-Channel (S2C) operation [52] to reshape $\mathbf{V}_U$ into the unified BEV space, $\mathbf{B}_U \in \mathbb{R}^{X \times Y \times (Z \times C)}$, effectively preserving semantic information while reducing computational cost. To recover the 3D volume features, we apply the inverse Channel-to-Spatial (C2S) reshaping to the BEV features. The unified BEV features serve as the target for the generation model, while the voxel features act as the input for the rendering decoder.

Multimodal Rendering Decoder. When given the unified volumetric features, the multimodal rendering decoder aims to decode multimodal sensor data, which can be divided into two distinct components: image reconstruction and LiDAR reconstruction. Practically, we represent a scene as an implicit signed distance function (SDF) field from Neus[46] as UniPAD [56] and use differentiable volume rendering to render multimodal data.

1) For image rendering, we sample camera rays from multi-view images $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ with the camera center $\mathbf{o}$ and viewing direction $\mathbf{d}$. For each ray $\mathbf{r}_i$, we sample N points $\{\mathbf{p}_i = (x_i, y_i, z_i)\}_{i=1}^{N}$ along the ray. For each sampled point $\mathbf{p}_i$, the corresponding features $\mathbf{v}_i$ are obtained from the voxel features $\mathbf{V}_U$ according to its position by trilinear interpolation. Then, the SDF value s_i is predicted by $\phi_{\mathrm{SDF}}(\mathbf{p}_i, \mathbf{v}_i)$, where ϕ_{SDF} represents a shallow MLP. Then, we render the camera feature descriptor by integrating the sampled features along rays:

$$\mathbf{F}_c = \sum_{i=1}^{N} w_i \mathbf{v_i}, w_i = \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \alpha_i = \max\left(\frac{\sigma_s(s_i) - \sigma_s(s_{i+1})}{\sigma_s(s_i)}, 0\right), \tag{1}$$

where $\sigma_s(x) = (1 + e^{-sx})^{-1}$ is a sigmoid function modulated by a learnable parameter s. After obtaining the 2D feature map $\mathbf{F}_c \in \mathbb{R}^{H_f^c \times W_f^c \times C_f}$ for each camera, we design a feature decoder to map the rendered features to the RGB image $\mathbf{I}_c \in \mathbb{R}^{H^c \times W^c \times 3}$ with enhanced high-frequency details. We employ MSE and LPIPS loss for rendered image supervision, *i.e.*, $\mathcal{L}_{\text{Camera}} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{lpips}}$.

2) For LiDAR rendering, following LiDAR-NeRF [43], we treat the oriented LiDAR laser beams as a set of camera rays. Slightly abusing the notation, let $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ be a ray casted from the LiDAR sensor, where $\mathbf{o}$ denotes the LiDAR center, and $\mathbf{d}$ represents the normalized direction vector of the corresponding beam. Then, similar to the camera rendering, we sample the ray points and get the corresponding features, and render the LiDAR depth measurement and feature descriptor as:

$$D = \sum_{i=1}^{N} w_i t_i, \mathbf{F}_l = \sum_{i=1}^{N} w_i \mathbf{v_i}. \tag{2}$$

The rendered depth D can convert into LiDAR points as $(x, y, z) = (D\cos(\alpha)\cos(\beta), D\cos(\alpha)\sin(\beta), D\sin(\alpha))$, where α is the vertical rotation and β is the horizontal rotation of viewing direction $\mathbf{d}$. The rendered LiDAR feature map is also decoded through a feature decoder to the view-dependent features of LiDAR, $\mathbf{I}_l \in \mathbb{R}^{H^l \times W^l \times 2}$, including the intensities and ray-drop probabilities. We employ L1 loss for depth optimization and MSE loss for intensities and ray-drop supervision, $\mathcal{L}_{\text{LiDAR}} = \mathcal{L}_{\text{depth}} + \mathcal{L}_i + \mathcal{L}_r$.

Compared to traditional VAEs [20], the rendering decoder naturally offers multiple benefits from its generalizable reconstruction capability. For instance, autonomous driving scenes typically involve multiple surround-view images. Previous works process each view independently using VAEs and apply attention modules as soft constraints to encourage cross-view consistency. However, this approach failed to maintain strict geometric consistency across different views. In contrast, the rendering decoder inherently supports multi-view consistent rendering by leveraging unified BEV features as a global scene constraint. Similarly, multimodal sensor data rendered from the shared BEV features also maintains consistency across different modalities. Furthermore, the rendering decoder allows for flexible adjustment of sensor parameters, *i.e.*, both intrinsic and extrinsic, during the rendering process. This enables intuitive camera control, which previously required complex designs and specialized training to achieve [27, 59].

Vector Quantization. To better leverage the diffusion model for generating BEV features, we further project the unified BEV features into a tokenized discrete space as VQ-VAE [44], which consists of three modules: BEV Patch Embedder, Vector Quantization, and BEV Feature Decoder.

1) BEV Patch Embedder. We firstly patchify the BEV features $\mathbf{B}_U \in \mathbb{R}^{X \times Y \times C}$ into a sequence of BEV patches $\{\mathbf{P}^i \in \mathbb{R}^{P \times P \times C}\}_{i=1}^{M}$, where P is the patch size, and $M = H^b W^b / P^2$ is the patch number. Then each BEV patch is further embedded to $\mathbf{z}_c \in \mathbb{R}^E$, where E is the embedded dimension. 2) Vector Quantization (VQ). We then define a discrete latent space $\{\mathbf{v}_1, \dots, \mathbf{v}_k, \dots, \mathbf{v}_K\} \in \mathbb{R}^{K \times E}$ as our codebook embedding, where K represents the maximum number of the embeddings. Taking the continuous latent vector $\mathbf{z}_c$ from the patch embedder, the VQ module outputs discrete latent vector $\mathbf{z}_d$ through the nearest neighbor search in the codebook. 3) BEV Feature Decoder. We finally feed the discrete BEV embeddings $\{\mathbf{z}_d^i\}_{i=1}^{M}$ to our BEV feature decoder by reshaping them into a grid format and then reconstructing the original BEV features. Detailed architecture is present in ?? of the supplementary material.

The overall VQ training loss $\mathcal{L}_{\text{vq}}$ includes the codebook loss $\mathcal{L}_{\text{code}}$ and the reconstruction loss $\mathcal{L}_{\text{re}}$. Due to the non-differentiable vector quantization operation, the codebook loss is defined as:

$$\mathcal{L}_{\text{code}} = \frac{1}{M} \sum_{i=1}^{M} \left(\left\| \mathbf{z}_d^i - \text{sg}(\ell_2(\mathbf{z}_c^i)) \right\|_2^2 + \left\| \text{sg}(\mathbf{z}_d^i) - \ell_2(\mathbf{z}_c^i) \right\|_2^2 \right), \tag{3}$$

where ℓ_2 means L2 normalization and sg denotes stop-gradient. We utilize MSE for the reconstruction loss of BEV features, and the final loss is defined as:

$$\mathcal{L}_{\text{vq}} = \mathcal{L}_{\text{re}} + \mathcal{L}_{\text{code}}. \tag{4}$$

Overall Optimization. The overall optimization target of our UAE is formulated as:

$$\mathcal{L}_{\text{UAE}} = \mathcal{L}_{\text{Camera}} + \mathcal{L}_{\text{LiDAR}} + \mathcal{L}_{\text{vq}}. \tag{5}$$

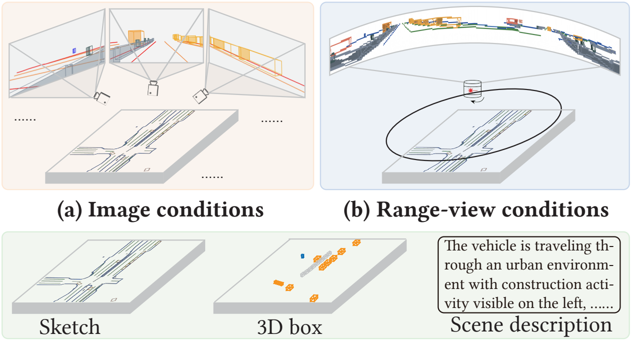
3.3 Unified LiDAR-Camera Generation

Latent BEV Feature DiT. Latent Diffusion Models (LDMs) [41] perform diffusion process on the latent space, using a pre-trained VAE to map between the input and the latent space. It iteratively denoises from a random Gaussian noise $\mathbf{z}_R^c$ for R steps with a denoiser $\mathcal{G}^c$ into a clean image latent $\mathbf{z}_0^c$. This VAE+diffusion formulation is widely adopted in image generation, and we also adopt this manner for our unified BEV features generation. As our BEV features already leverage vector quantization into discrete space, we do not need to apply additional VAE mapping. Given scene conditions $\mathbf{S}$, the goal is to generate corresponding BEV features from latent variables $\epsilon \sim \mathcal{N}(0, I)$, *i.e.*, $\mathbf{B}_U = \mathcal{G}(\mathbf{S}, \epsilon\}$. Then the generated BEV features are decoded to multimodal sensor data through the multimodal rendering decoder of our UAE. To enhance the quality of generated BEV features, we adapt the advanced DiT [65] as our denoiser $\mathcal{G}$ and apply a cross-attention mechanism to integrate scene conditions $\mathbf{S}$, *i.e.*, textual descriptions, road sketches, and 3D boxes. Denoting $z_\tau^b(\epsilon) = \sqrt{\bar{\alpha}_\tau} z_0^b + \sqrt{1 - \bar{\alpha}_\tau}\epsilon$ as noisy latent, where τ is a timestep, $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise, $\bar{\alpha}_\tau$ is hyper-parameter, the diffusion process is:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0^b, \epsilon, \tau, \mathbf{S}} \left[\| \epsilon - \mathcal{G}_\theta^b(z_\tau^b(\epsilon), \tau, \mathbf{S}) \|_2^2 \right]. \tag{6}$$

Specifically, as illustrated in Fig. 2, we incorporate the ControlNet [62] branch into the DiT model to enable road sketch conditions. Inspired by PixArt-δ [6], we create a trainable copy of the first 13 blocks of the model. These duplicated blocks are integrated with the corresponding base blocks through a learnable zero linear layer. Each duplicated block combines the road sketch features, ensuring precise control with the provided sketch conditions.

Scene Conditions Encoding. To describe a driving scenario, we adopt comprehensive scene conditions outlined in [9, 33]. As illustrated in Fig. 3, unlike previous approaches that rely on modality-specific conditions, our method adopts unified conditions to generate aligned multimodal sensor data. Specifically, scene conditions $\mathbf{S} = \{\mathbf{M}, \mathbf{B}, \mathbf{T}\}$ include a road sketch $\mathbf{M} \in \{0, 1\}^{w \times h \times c}$ representing a $w \times h$ meter road area in BEV with c semantic classes, 3D bounding boxes $\mathbf{B} = \{(\mathbf{b}_i, \mathbf{h}_i, \mathbf{l}_i, \mathbf{c}_i)\}_{i=1}^{n}$ where each object is described by a box $\mathbf{b}_i = \{(x_j, y_j, z_j)\}_{j=1}^{8} \in \mathbb{R}^{8 \times 3}$, heading $\mathbf{h}_i \in [-180, 180)^{n \times 1}$, instance


(a) Image conditions

(b) Range-view conditions

......
......
......
The vehicle is traveling through an urban environment with construction activity visible on the left, ...
Sketch
3D box
Scene description

(c) Unified conditions (ours)

Figure 3: Previous modality-specific conditions and our unified conditions.

id $\mathbf{l}_i \in [0,1]^{n\times 1}$, and caption $\mathbf{c}_i \in \mathcal{C}$, textual descriptions $\mathbf{T}$ summarying information for the whole scene (*e.g.*, weather and time of day). The layout entries, *i.e.*, instance details such as box coordinates, heading, and ID, are encoded by the Fourier Embedder [34], F, and then are concatenated and processed through an MLP into a unified embedding. The textual input is encoded into 200 tokens using the T5 [39] language model, E_{T5}. The road sketches are extracted latent features by a pre-trained VAE, E_{VAE}. The encoding of unified scene conditions can be formulated as:

$$\begin{aligned} \mathbf{B}' &= \text{MLP}(F(\mathbf{b}) + F(\mathbf{h}) + F(\mathbf{l}) + E_{\text{T5}}(\mathbf{c})), \\ \mathbf{M}' &= E_{\text{VAE}}(\mathbf{M}),\ \mathbf{T}' = E_{\text{T5}}(\mathbf{T}). \end{aligned} \tag{7}$$

We incorporate these condition embeddings into the DiT model through cross-attention mechanisms, facilitating flexible and fine-grained control:

$$\begin{aligned} &\mathbf{q} = \text{MLP}(\mathbf{z}^b_{\text{in}}),\ \mathbf{k} = \text{MLP}([\mathbf{B}', \mathbf{T}']),\ \mathbf{v} = \text{MLP}([\mathbf{B}', \mathbf{T}']), \\ &\text{CA}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{Softmax}(\frac{\mathbf{q}\cdot\mathbf{k}^T}{\sqrt{d}})\cdot\mathbf{v}. \end{aligned} \tag{8}$$

The road sketches feature is integrated in the duplicated blocks as:

$$\mathbf{z}^b_{\text{out}} = \text{DiT}(\mathbf{z}^b_{\text{in}}) + \texttt{Zero}(\text{Control}(\mathbf{z}^b_{\text{in}} + \mathbf{M}')), \tag{9}$$

where Zero denotes the learnable zero linear layer.

Optimization. With the latest advancement on LDMs [65], we replace IDDPM [36] with rectified flow [28] for increased stability and reduced inference steps. Rectified flow defines the forward process between data and normal distributions as $z^b_\tau = (1-\tau)z^b_0 + \tau z^b_1$, and the loss function in Eq. (6) is rewritten as:

$$\mathcal{L}_{\text{rf}} = \mathbb{E}_{z^b_1, z^b_0, \tau, \mathbf{S}}\left[\|\mathcal{G}^b_\theta(z^b_\tau, \tau, \mathbf{S}) - (z^b_1 - z^b_0)\|^2_2\right]. \tag{10}$$

Moreover, we extend the Classifier-free Guidance (CFG) [13] strategy from the text condition to 3D boxes and road sketches to enhance control precision and visual quality. CFG aims to enhance the alignment between generated images and specified conditions, which simultaneously performs both conditional and unconditional denoising during training and combines the two estimated scores during inference. In practice, we randomly set each condition to a null ϕ with a 5% probability during training. The guidance scale $\lambda_M, \lambda_B, \lambda_T$ controls the alignment between the sampling results and the conditions. Drawing inspiration from IP2P [1], we apply the unconditional denoising results to each condition individually, which can be formulated as:

$$\begin{aligned} \tilde{\mathcal{G}}_\theta(z^b_\tau, B, M, T) = \ &\mathcal{G}_\theta(z^b_\tau, \phi, \phi, \phi) \\ &+ \lambda_T\cdot(\mathcal{G}_\theta(z^b_\tau, \phi, \phi, T) - \mathcal{G}_\theta(z^b_\tau, \phi, \phi, \phi)) \\ &+ \lambda_M\cdot(\mathcal{G}_\theta(z^b_\tau, \phi, M, T) - \mathcal{G}_\theta(z^b_\tau, \phi, \phi, T)) \\ &+ \lambda_B\cdot(\mathcal{G}_\theta(z^b_\tau, B, M, T) - \mathcal{G}_\theta(z^b_\tau, \phi, M, T)). \end{aligned} \tag{11}$$

Table 1: Multimodal sensor generation results.

Method	Tokenizer	*Camera Generation*		
		FID↓	CLIP↑	mAP↑
Single-Modality				
BEVGen [42]	VQ-VAE	25.54	71.23	–
BEVControl [57]	VQ-VAE	24.85	82.70	19.64
DriveDreamer [47]	VQ-VAE	52.60	–	–
DriveDreamer-2 [64]	VQ-VAE	25.00	–	–
WoVoGen [31]	VQ-VAE	27.60	–	–
MagicDrive [9]	VQ-VAE	16.20	82.47	12.30
Panacea [49]	VQ-VAE	16.96	84.23	–
Drive-WM [48]	VQ-VAE	15.80	–	20.66
MagicDriveDiT [11]	VQ-VAE	20.91	–	17.65
OmniGen	**UAE-C**	22.15	82.76	19.57
Unified				
OmniGen	**UAE-LC**	21.01	83.54	20.41

Method	Tokenizer	*LiDAR Generation*		
		FRD↓	MMD↓	JSD↓
Single-Modality				
LiDARVAE [3]	2DGrid-VAE	–	11.0	-
LiDARGen [68]	N/A	–	19.0	0.160
RangeLDM [15]	Range-VAE	492.49	2.75	0.054
LidarDM [69]	SDF-VAE	–	3.51	0.118
OmniGen	**UAE-L**	562.89	3.17	0.117
Unified				
OmniGen	**UAE-LC**	519.73	2.94	0.105

MMD has been multiplied by 10^4.

"LC" represents the LiDAR and Camera fusion.

4 Experiment

Our experiments are conducted on the popular NuScenes dataset [4]. For additional details, including dataset, metrics, baselines, and implementations, please see ?? in the supplementary material.

4.1 Main Results

Unified multimodal sensor generation. As shown in Table 1, we compare our model with specialized single-modality generation methods. Although our method does not achieve state-of-the-art (SOTA) performance for every metric, OmniGen, as a unified multimodal framework, achieves comparable or even superior quality in camera and LiDAR sensor generation. Specifically, OmniGen achieves 21.01 FID and 20.41 mAP for generated camera data,

Table 2: Generalizable multimodal reconstruction results.

Camera Recon	Train set		Val set	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SelfOcc [18]	20.67	0.556	–	–
UniPAD [16]	19.44	0.497	–	–
DistillNeRF [45]	28.01	0.872	–	–
UAE-C	30.29	0.908	30.13	0.903
UAE-LC	**30.45**	**0.913**	**30.21**	**0.909**

LiDAR Recon	Train set		Val set	
	Chamfer↓	F-score↑	Chamfer↓	F-score↑
UAE-L	0.869	0.734	1.068	0.713
UAE-LC	**0.634**	**0.763**	**0.793**	**0.742**

Table 3: Generation data augmentation for perception.

Method	Modality	mAP↑	NDS↑
BEVFormer [24]	C	25.2	35.4
+ **OmniGen**	C	**27.1** (+1.9)	**37.1** (+1.7)
BEVFusion [29]	LC	68.5	71.4
+ **OmniGen**	LC	**70.1** (+1.6)	**72.8** (+1.4)

Table 4: Generation data augmentation for planning.

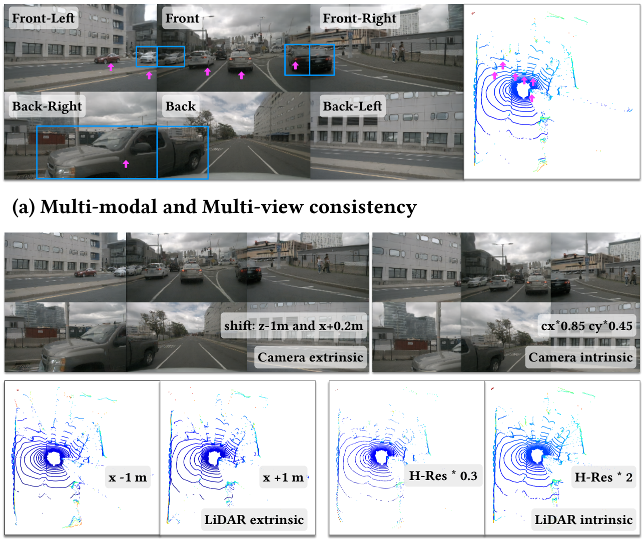
Method	Modality	Avg. L2 (m) ↓	Avg. Collision (%) ↓
UniAD [16]	C	1.03	0.31
+ **OmniGen**	C	**0.99** (+3.9%)	**0.29** (+6.4%)
FusionAD [60]	LC	0.81	0.12
+ **OmniGen**	LC	**0.77** (+4.9%)	**0.11** (+8.3%)

and 2.94×10^{-4} MMD and 0.105 JSD for LiDAR data. Furthermore, thanks to our independent BEV encoders for each modality, our UAE (*i.e.*, UAE-LC) also supports single-modality operation, *i.e.*, UAE-C and UAE-L. Experimental results show that multimodal generation consistently outperforms our single-modality generation, further demonstrating the effectiveness of our unified model. We hope our framework can inspire the community and foster collaborative efforts to improve the results to SOTA together.

Generalizable LiDAR-camera reconstruction. As shown in Table 2, our UAE achieves state-of-the-art performance in generalizable LiDAR-camera multimodal reconstruction. Specifically, UAE significantly outperforms the previous best generalizable single-camera modality method, DistillNeRF [45], with a notable improvement (*i.e.*, +2.44 PSNR). Moreover, our UAE-L and UAE-LC first achieves generalizable LiDAR reconstruction with the chamfer distance of 0.869 and 0.634, respectively. Additionally, multimodal reconstruction also surpasses single-modality reconstruction, further highlighting the advantages of our unified framework.

Downstream tasks. As shown in Table 3 and Table 4, we utilize our OmniGen to produce augmented data with corresponding conditions, aiming to enhance downstream tasks. As illustrated by the improvements, our OmniGen effectively generates multimodal sensor data, facilitating enhanced perception and planning in autonomous driving. Specifically, OmniGen boosts multimodal BEVFusion with +1.6 mAP and FusionAD with +4.9% L2 metric.

4.2 Qualitative Results



(a) Multi-modal and Multi-view consistency

(b) Sensor extrinsic and intrinsic controllability

Figure 4: Qualitative results for multimodal sensor generation. H-res denotes the horizontal resolution.

Multi-modal and multi-view consistency. As shown in Fig. 4 (a), our OmniGen exhibits excellent multi-modal and multi-view consistency, as highlighted by the blue boxes and pink arrows. This demonstrates the effectiveness of our unified method, which leverages unified BEV features as a global scene constraint.

Sensor intrinsic and extrinsic controllability. As shown in Fig. 4 (b), UAE enables the flexible adjustment of sensor parameters, including both intrinsic and extrinsic settings, during the rendering process, which showcases the superior design of our framework.

4.3 Ablation Study

Ablations on UAE. As shown in Table 5 and Fig. 5, we investigate various designs of the render decoder in UAE. For A1-MLP, we introduce additional MLPs to further extract voxel features; however, this does not improve the results and instead increases the number of parameters. For A2-w/o-decoder, we use additional MLPs to directly output RGB values, removing the feature decoder altogether. This leads to a significant performance drop, highlighting the importance of the feature decoder. Additionally, we ablate different representations for rendering, including NeRF and SDF, both of which produce comparable results. The SDF performs slightly better and is therefore adopted in the final design.

Table 5: Ablation on UAE.

Module	PSNR↑	SSIM↑	LPIPS↓
Architecture			
A1-MLP Fig5 (b)	29.42	0.878	0.049
A2-w/o-decoder Fig5 (c)	26.61	0.802	0.212
UAE Fig5 (a)	**30.21**	**0.909**	**0.033**
Render Representation			
NeRF Fig5 (d)	29.97	0.891	0.069
SDF Fig5 (a)	**30.21**	**0.909**	**0.033**

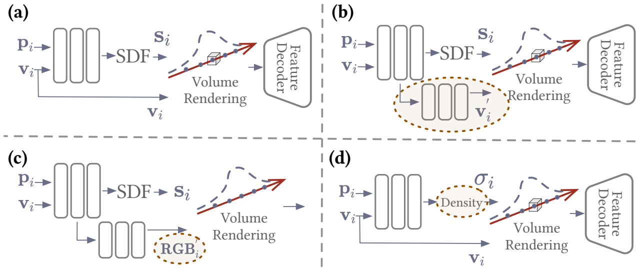


Figure 5: Different architecture designs of UAE.

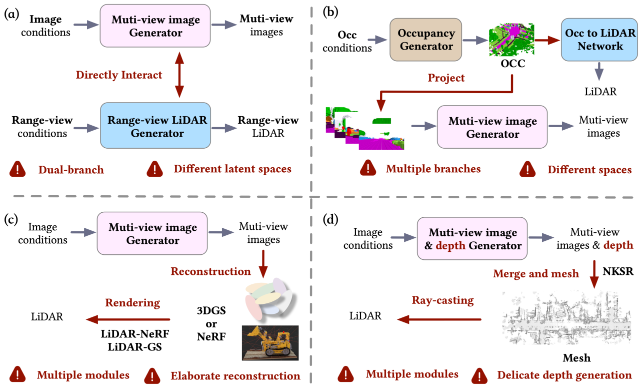


Figure 6: Multimodal sensor generation solutions (zoom-in for better views).

5 Discussion

In this section, we aim to provide valuable insights for the unified multimodal generation field. More discussions about Future Work of Multimodal Generation, Limitations and Unsuccessful Attempts, please see ?? in the supplementary material.

5.1 Mutimodal Sensor Generation Solutions

During the early design phase of OmniGen, we explored various alternative frameworks. Although these ideas were not pursued further due to certain limitations, *e.g.*, not unified, we present them here to share potential ideas for future research. As shown in Fig. 6, these frameworks build upon the camera generation pipeline as the primary branch, leveraging its superior camera generation performance, while incorporating the LiDAR modality in different ways.

- a) A dual-branch design, where each modality operates in its own feature space and cross-modal alignment is enforced by attention mechanisms. Due to the inherent differences in representation spaces, achieving alignment proves challenging, which leads to limited performance potential for the overall framework.
- b) Some studies [22, 32, 54] use semantic occupancy as an intermediate representation to improve generation quality. Based on this approach, we explored adding an occupancy-to-LiDAR branch. However, this results in a complex framework with multiple branches, and its performance is restricted by limitations in the occupancy generation process, such as resolution.
- c) Expanding upon the existing camera branch, we incorporated a reconstruction stage using NeRF or 3DGS to reconstruct the scene, followed by rendering LiDAR data through methods such as [7, 43]. However, this framework suffers from inefficiencies due to its multi-stage nature, particularly the time-consuming and labor-intensive reconstruction process.
- d) Building on the existing camera branch, we integrated it with per-view depth map generation and then merged the generated depth maps into a mesh using methods such as NKSR [17]. Subsequently, LiDAR data can be synthesized through ray-casting. However, this framework is limited by its multi-stage complexity and the challenges of generating precise depth.

6 Conclusion

In this paper, we introduced OmniGen, a unified multimodal sensor generation framework for autonomous driving that enables the unified generation of aligned LiDAR and camera data. Our approach addresses the limitations of existing single-modality generation methods by establishing a unified BEV-based representation space, proposing UAE, a generalizable multimodal reconstruction for multimodal autoencoder, and incorporating a ControlNet-Transformer model to synthesize multimodal sensor data under flexible conditions. Extensive experiments demonstrate that our framework not only achieves state-of-the-art performance in multimodal reconstruction but also generates LiDAR and camera data with cross-modality alignment and flexible sensor control. These advancements enhance the quality and usability of synthetic sensor data, further benefiting downstream tasks such as perception and planning in autonomous driving. Moreover, we provide valuable insights into the designs of the unified multimodal generation framework, hoping to inspire future research into more efficient multimodal generation.

7 Acknowledgments

This work is supported by National Key Research and Development Program of China (2024YFE0203100), National Natural Science Foundation of China (NSFC) under Grants No.62476293, Shenzhen Science and Technology Program No.GJHZ20220913142600001, Nansha Key R&D Program under Grant No.2022ZD014, and General Embodied AI Center of Sun Yat-sen University.

References

[1] 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*, pages 18392–18402, 2023.

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3:1, 2024.

[3] Lucas Caccia, Herke Van Hoof, Aaron Courville, and Joelle Pineau. Deep generative modeling of lidar data. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5034–5040. IEEE, 2019.

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021.

[6] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-δ: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024.

[7] Qifeng Chen, Sheng Yang, Sicong Du, Tao Tang, Peng Chen, and Yuchi Huo. Lidar-gs: Real-time lidar re-simulation using gaussian splatting. *arXiv preprint arXiv:2410.05111*, 2024.

[8] Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. *arXiv preprint arXiv:2412.04842*, 2024.

[9] 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.

[10] 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.

[11] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrivedit: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024.

[12] 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.

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[15] Qianjiang Hu, Zhimin Zhang, and Wei Hu. Rangeldm: Fast realistic lidar point cloud generation. *arXiv preprint arXiv:2403.10094*, 2024.

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[17] Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural lidar fields for novel view synthesis. *ICCV*, 2023.

[18] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19946–19956, 2024.

[19] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.

[20] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[21] Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. *arXiv preprint arXiv:2412.05435*, 2024.

[22] Leheng Li, Weichao Qiu, Yingjie Cai, Xu Yan, Qing Lian, Bingbing Liu, and Ying-Cong Chen. Syntheocc: Synthesize geometric-controlled street view images through 3d semantic mpis. *arXiv preprint arXiv:2410.00337*, 2024.

[23] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. In *Advances in Neural Information Processing Systems*, 2022.

[24] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[25] 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. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022.

[26] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[27] 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. *arXiv preprint arXiv:2407.19468*, 2024.

[28] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

[29] 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 *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023.

[30] Hannan Lu, Xiaohe Wu, Shudong Wang, Xiameng Qin, Xinyu Zhang, Junyu Han, Wangmeng Zuo, and Ji Tao. Seeing beyond views: Multi-view driving scene video generation with holistic attention. *arXiv preprint arXiv:2412.03520*, 2024.

[31] 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.

[32] 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.

[33] Enhui Ma, Lijun Zhou, Tao Tang, Zhan Zhang, Dong Han, Junpeng Jiang, Kun Zhan, Peng Jia, Xianpeng Lang, Haiyang Sun, et al. Unleashing generalization of end-to-end autonomous driving with controllable long video generation. *arXiv preprint arXiv:2406.01349*, 2024.

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.

[35] Kazuto Nakashima and Ryo Kurazume. Lidar data synthesis with denoising diffusion probabilistic models. *arXiv preprint arXiv:2309.09256*, 2023.

[36] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021.

[37] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[38] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[40] 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*, pages 14738–14748, 2024.

[41] 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*, pages 10684–10695, 2022.

[42] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird’s-eye view layout. *arXiv preprint arXiv:2301.04634*, 2023.

[43] Tang Tao, Longfei Gao, Guangrun Wang, Yixing Lao, Peng Chen, Hengshuang Zhao, Dayang Hao, Xiaodan Liang, Mathieu Salzmann, and Kaicheng Yu. Lidar-nerf: Novel lidar view synthesis via neural radiance fields. *arXiv preprint arXiv:2304.10406*, 2023.

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[45] Letian Wang, Seung Wook Kim, Jiawei Yang, Cunjun Yu, Boris Ivanovic, Steven L Waslander, Yue Wang, Sanja Fidler, Marco Pavone, and Peter Karkus. Distillnerf: Perceiving 3d scenes from single-glance images by distilling neural fields and foundation model features. *arXiv preprint arXiv:2406.12095*, 2024.

[46] 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.

[47] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

[48] 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. *arXiv preprint arXiv:2311.17918*, 2023.

[49] 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. *arXiv preprint arXiv:2311.16813*, 2023.

[50] Yang Wu, Kaihua Zhang, Jianjun Qian, Jin Xie, and Jian Yang. Text2lidar: Text-guided lidar point cloud generation via equirectangular transformer. *arXiv preprint arXiv:2407.19628*, 2024.

[51] Bin Xie, Yingfei Liu, Tiancai Wang, Jiale Cao, and Xiangyu Zhang. Glad: A streaming scene generator for autonomous driving. *arXiv preprint arXiv:2503.00045*, 2025.

[52] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. M2bev: multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022.

[53] Yuwen Xiong, Wei-Chiu Ma, Jingkang Wang, and Raquel Urtasun. Learning compact representations for lidar completion and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1074–1083, 2023.

[54] Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng-zhong Xu, and Jianbing Shen. Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. *arXiv preprint arXiv:2411.11252*, 2024.

[55] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[56] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15238–15250, 2024.

[57] 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.

[58] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[59] Yining Yao, Xi Guo, Chenjing Ding, and Wei Wu. Mygo: Consistent and controllable multi-view driving video generation with camera control. *arXiv preprint arXiv:2409.06189*, 2024.

[60] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, et al. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. *arXiv preprint arXiv:2308.01006*, 2023.

[61] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021.

[62] 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*, pages 3836–3847, 2023.

[63] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5449–5458, 2022.

[64] 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.

[65] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[66] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018.

[67] Yingshuang Zou, Yikang Ding, Chuanrui Zhang, Jiazhe Guo, Bohan Li, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, and Haoqian Wang. Mudg: Taming multi-modal diffusion with gaussian splatting for urban scene reconstruction. *arXiv preprint arXiv:2503.10604*, 2025.

[68] Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point clouds. In *European Conference on Computer Vision*, pages 17–35. Springer, 2022.

[69] Vlas Zyrianov, Henry Che, Zhijian Liu, and Shenlong Wang. Lidardm: Generative lidar simulation in a generated world. *arXiv preprint arXiv:2404.02903*, 2024.