SCALABLE AND GENERALIZABLE AUTONOMOUS DRIV-ING SCENE SYNTHESIS

Anonymous authors

000

001

002003004

006

008 009

010 011

012

013

014

015

016

017

018

019

021

023

024

025

026027028

029

031

033

037

038

040

041 042

043

044

046

047

048

049

051

052

Paper under double-blind review

ABSTRACT

Generative modeling has shown remarkable success in vision and language, inspiring research on synthesizing autonomous driving scenes. Existing multi-view synthesis approaches commonly operate in image latent spaces with cross-attention to enforce spatial consistency, but they are tightly bound to camera configurations, which limits dataset scalability and model generalization. We propose BEV-VAE, a variational autoencoder that unifies multi-view images into a compact bird's-eyeview (BEV) representation, enabling encoding from arbitrary camera layouts and decoding to any desired viewpoint. Through multi-view image reconstruction and novel view synthesis, we show that BEV-VAE effectively fuses multi-view information and accurately models spatial structure. This capability allows it to generalize across camera configurations and facilitates scalable training on diverse datasets. Within the latent space of BEV-VAE, a Diffusion Transformer (DiT) generates BEV representations conditioned on 3D object layouts, enabling multi-view image synthesis with enhanced spatial consistency on nuScenes and achieving the first complete seven-view synthesis on AV2. Finally, synthesized imagery significantly improves the perception performance of BEVFormer, highlighting the utility of scalable and generalizable scene synthesis for autonomous driving.



Figure 1: **Autonomous driving scene synthesis from AV2 to nuScenes.** BEV-VAE with DiT generates a BEV representation from 3D bounding boxes of AV2, which can then be decoded into multi-view images according to the camera configurations of nuScenes.

1 Introduction

The significant impact of generative modeling on vision (Rombach et al., 2022) and language (Achiam et al., 2023) has motivated research on the synthesis of autonomous driving scenes. Specifically, multi-view image synthesis conditioned on 3D object annotations can vary both object appearance and scene background while preserving the ground-truth 3D box locations. This enables 3D perception models (Li et al., 2024b) to learn the correspondence between changing visual appearance and fixed spatial positions. However, the effectiveness of such synthesized imagery critically depends on both per-view quality and cross-view consistency. Existing approaches (Li et al., 2024a; Gao et al., 2023; Wen et al., 2024; Wang et al., 2024) typically achieve multi-view synthesis by training generative models in the image latent space, ensuring spatial consistency through cross-view attention. Although this paradigm can ensure consistency, it introduces significant computational costs and high modeling complexity. Moreover, it is inherently tied to specific vehicle types and camera layouts, limiting both the scale of available training data and the generalizability of the synthesized imagery. For example, a model trained on seven camera views cannot be applied directly to vehicles equipped with six.

•	Dataset	# Frames	# Cameras	# Classes	Recording Locations
	WS101	17K	5	0	London, San Francisco Bay Area
	nuScenes	155K	6	23	Boston, Pittsburgh, Las Vegas, Singapore
	AV2	224K	7	30	Austin, Detroit, Miami, Pittsburgh, Palo Alto, Washington DC
	nuPlan	3.11M	8	7	Boston, Pittsburgh, Las Vegas, Singapore

Table 1: Comparison of autonomous driving datasets with full 360° multi-camera coverage. These datasets vary in dataset scale, camera configurations, 3D annotation categories, and recording locations, where WS101 does not provide 3D annotations.

In reality, multi-view images with varying camera layouts are only different projections of the same scene. Motivated by this insight, we introduce BEV-VAE, a variational autoencoder that unifies multi-view images into a compact BEV representation and utilizes this latent space for generative modeling. The BEV representation integrates the semantics of all views and constructs the 3D structure of the scene, enabling encoding from arbitrary camera layouts and decoding to any desired viewpoints. It avoids explicitly modeling spatial relationships across views, which substantially reduces computational cost and modeling complexity for generative modeling. In addition, training can be performed on multiple datasets that cover different types of vehicle and camera layouts. This overcomes the data isolation limitations of existing methods and enables scalable and generalizable autonomous driving scene synthesis across datasets and viewpoints.

We systematically evaluate the scalability and generalizability of BEV-VAE across four autonomous driving datasets (Zürn et al., 2024; Caesar et al., 2020; Wilson et al., 2023; Caesar et al., 2021), which vary in dataset scale, camera configurations and recording locations (see Tab. 1). The spatial modeling capability of BEV-VAE is validated by multi-view image reconstruction, as the reconstruction fidelity reflects its ability to construct the spatial relationships between objects and the background in the scene. Novel view synthesis is further achieved by modifying camera poses when decoding the BEV representation into images, directly demonstrating that BEV-VAE encodes precise spatial structure and comprehensive scene semantics. In addition, BEV-VAE overcomes the data isolation caused by varying vehicle camera setups, effectively integrating datasets collected worldwide and greatly increasing the diversity of training data. Models trained on mixed datasets achieve significantly higher reconstruction quality than trained individually on AV2 or nuScenes, demonstrating the scalability of BEV-VAE. Meanwhile, multi-dataset joint training enables BEV-VAE to generalize across different vehicle types and camera setups. For example, it can convert images from the 8-camera configuration of nuPlan to the 7-camera setup of AV2 or the 6-camera setup of nuScenes. This indicates that BEV-VAE generalizes not only across camera poses but also camera intrinsics. Furthermore, BEV-VAE demonstrates zero-shot multi-view reconstruction on WS101, enabled by training on datasets with diverse camera configurations, and achieves further enhanced reconstruction quality through fine-tuning on its specific setups.

We train a Diffusion Transformer (DiT) (Peebles & Xie, 2023) in the latent space of BEV-VAE to enable multi-view image synthesis conditioned on 3D object layouts. These object layouts are encoded as occupancy grids that are spatially aligned with the BEV representation, allowing precise specification of object positions and counts in the scene, analogous to ControlNet (Zhang et al., 2023). Specifically, we achieve multi-view image synthesis with enhanced spatial consistency on nuScenes, and are the first to synthesize images for all seven camera views on AV2. Moreover, by decoding the BEV representation with the camera configuration of nuScenes, the seven images synthesized on AV2 can be directly converted into the six corresponding views of nuScenes. Finally, we show that synthesized imagery can significantly improve the performance of BEVFormer on nuScenes, validating the effectiveness of synthesis-based appearance diversification as a data augmentation strategy for perception.

2 RELATED WORK

2.1 BIRD'S-EYE-VIEW REPRESENTATION

Autonomous driving relies on Bird's Eye View (BEV) to integrate information from multiple camera perspectives. The construction of BEV representations is typically approached in two ways: bottom-up and top-down. Bottom-up methods (Philion & Fidler, 2020; Huang et al., 2021; Liu et al., 2023) estimate depth to lift 2D features into 3D space before fusing them into BEV. In contrast, top-down methods (Li et al., 2024b; Hu et al., 2023) employ deformable attention (DA) and query mechanisms

to efficiently aggregate features through dynamic sampling of key regions. These two paradigms provide complementary perspectives for designing BEV representations.

2.2 Variational AutoEncoder for Generative Modeling

VAE provides an efficient latent-variable framework for generative modeling. VQVAE (Van Den Oord et al., 2017) extends VAE by introducing a discrete codebook, which allows latent representations to be combined with Transformers for autoregressive image generation. Building on this, VQ-GAN (Esser et al., 2021) further improves image quality by incorporating adversarial and perceptual losses (Johnson et al., 2016), and ViT-VQGAN (Yu et al., 2021) leverages ViT (Dosovitskiy, 2020) in both encoder and decoder to enhance global context modeling and codebook utilization. In parallel, diffusion models (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022) achieve high-fidelity image generation through iterative denoising, with DiT (Peebles & Xie, 2023) combining diffusion and Transformer architectures to enable scalable generative modeling. Despite these advances, improved VAE variants remain important, as they provide higher compression ratios (Chen et al., 2024) or alignment with foundation model representations (Yao et al., 2025), supporting scalable and effective generative modeling.

2.3 AUTONOMOUS DRIVING SCENE SYNTHESIS

Autonomous driving scene synthesis is typically approached via multi-view synthesis, in which a 3D scene is inherently represented through multiple 2D images. BEVGen (Swerdlow et al., 2024) uses autoregressive generation to produce multi-view images based on BEV layouts. It constructs direction vectors for cameras and BEV layouts, maps them to the BEV ego-vehicle coordinate system via camera parameters, and integrates their inner product as an attention bias to enhance spatial consistency. However, subsequent works have increasingly adopted diffusion-based generation methods, fine-tuning Stable Diffusion to transfer its conditional generation capabilities to the autonomous driving domain. Driving Diffusion (Li et al., 2024a), Magicdrive (Gao et al., 2023), and Panacea (Wen et al., 2024) utilize cross-attention on adjacent view images to ensure consistency between perspectives. MagicDrive integrates camera pose information by encoding camera parameters similar to NeRF (Mildenhall et al., 2021), while Panacea extends this approach by generating pseudo-RGB images of camera frustum directions and embedding pose information through ControlNet (Zhang et al., 2023). Additionally, DriveWM (Wang et al., 2024) uses self-attention to fuse spatially aligned features across views and predicts stitched views between nonadjacent references to maintain multiview spatial consistency. However, existing methods generally underutilize the physical information of camera parameters and lack explicit 3D spatial modeling, restricting them to generating images from fixed viewpoints. This limitation confines scene modeling to a narrow set of perspectives and hinders cross-platform generalization in autonomous driving scenarios.

3 Method

3.1 Overall Architecture of BEV-VAE

BEV-VAE consists of a Transformer-based encoder E, decoder G, and a StyleGAN discriminator D. The encoder E maps multi-view images into a latent Gaussian distribution via its image, scene, and state encoders, from which state features are sampled via reparameterization. The decoder G, comprising state, scene, and image decoders, reconstructs spatially consistent multi-view images from the state features. The discriminator D distinguishes real from reconstructed images, guiding G with adversarial loss. Both encoder E and decoder G are trained with KL divergence, reconstruction, and adversarial losses.

3.1.1 ENCODER

Image Encoder employs ViT with a patch size of 8 to encode a 256×256 image into a 32×32 token sequence. To capture semantic information and local details for 3D scene encoding, an upsampling-only FPN Lin et al. (2017) constructs a three-level feature pyramid to enhance multi-scale representation. The process can be formulated as: $F_{img} = \text{FPN}(\mathbf{E}_{img}(x)) = \text{Concat}(F_{img}^0, F_{img}^1, F_{img}^2)$,

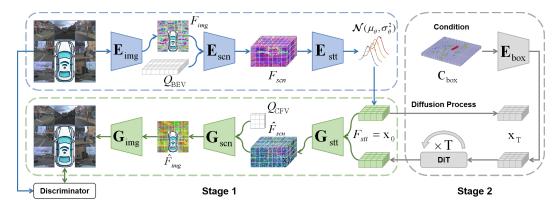


Figure 2: **Overall architecture of BEV-VAE with DiT for autonomous driving scene synthesis.** In Stage 1, BEV-VAE learns to encode multi-view images into a compact latent space in BEV and reconstruct them, modeling the spatial structure and representing the scene semantics. In Stage 2, DiT is trained with Classifier-Free Guidance (CFG) in this latent space to generate BEV representations from random noise, which are then decoded into multi-view images.

where $F^i_{img} \in \mathbb{R}^{V \times L_i \times C} (i \in [0,2])$ are the multi-scale flattened image features with C=96 and sequence length $L_i=32 \times 32 \times 2^{2i}$. Here, V is the number of views.

Scene Encoder utilizes a deformable attention mechanism to construct 3D scene features by extracting multiview image features. A 128×128 grid of pillars is pre-defined around the ego vehicle in BEV, each with a height of 8. All reference points in the same pillar share a learnable query, while different height positions are distinguished through positional encoding. The reference points of scene features are projected onto image features by camera parameters, enabling BEV queries to aggregate spatially aligned features from multiview image features via deformable attention. The process can be formulated as: $F_{scn} = \frac{1}{|\mathcal{V}_{hit}|} \sum_{v \in \mathcal{V}_{hit}} \mathrm{DA}(Q_{\mathrm{BEV}}, P_{\mathrm{BEV}}, F_{img}^{(v)})$, where $Q_{\mathrm{BEV}} \in \mathbb{R}^{L_Q \times C}$ are the flattened 3D BEV queries with C = 96, $P_{\mathrm{BEV}} \in \mathbb{R}^{L_Q \times 3}$ denote the corresponding reference points, $F_{img}^{(v)} \in \mathbb{R}^{L_V \times C}$ is the image feature sequence of the view v, and the set \mathcal{V}_{hit} refers to the views containing projected reference points, ensuring that only relevant views contribute to the aggregated scene feature. Here, $L_Q = 8 \times 128 \times 128$ is the BEV query sequence length, and $L_V = \sum_{i=0}^2 (32 \times 32 \times 2^{2i})$ is the total image feature sequence length across resolutions.

State Encoder integrates multi-height scene features in BEV by concatenating them along the height dimension, reshaping the input from $96 \times 8 \times 128 \times 128$ to $768 \times 128 \times 128$. It then partitions the features into 32×32 patches along the horizontal plane, reducing the computational cost while introducing local receptive fields. Finally, it applies self-attention to model global spatial relationships and encode highly compressed spatial state features.

3.1.2 Decoder

State Decoder is responsible for reconstructing structurally detailed 3D scene features from the compressed 2D state representation. It first applies self-attention to capture global spatial relationships, and then regroups the features to restore horizontal and height structures. The state features are first expanded from 32×32 to 128×128 along the horizontal plane through deconvolution, then further transformed from $768 \times 128 \times 128$ to the original multi-height format $96 \times 8 \times 128 \times 128$ through dimension partitioning. To refine 3D scene feature decoding, a downsampling-only FPN is employed, effectively reconstructing detailed structures across scales. The process can be formulated as: $\hat{F}_{scn} = \text{FPN}(\mathbf{G}_{\text{stt}}(\hat{x})) = \text{Concat}(\hat{F}_{scn}^0, \hat{F}_{scn}^1, \hat{F}_{scn}^2)$, where $\hat{F}_{scn}^i \in \mathbb{R}^{L_i \times C} (i \in [0, 2])$ are the reconstructed multi-scale flattened scene features with C = 96 and sequence length $L_i = 8 \times 128 \times 128 \times 2^{-3i}$.

Scene Decoder transforms scene features from the Bird's Eye View (BEV) to the Camera's Frustum View (CFV) and aggregates multi-depth information to reconstruct image features. A 32×32 frustum of rays is predefined per camera, each spanning 60 depth levels. All reference points along the same ray share a learnable query, while different depth positions are distinguished through positional encoding. Similar to the projection of reference points of scene features from BEV onto image features via camera parameters, reference points of scene features in CFV can also be

projected to BEV, enabling CFV queries to construct features along depth dimensions for different views via deformable attention. Furthermore, CFV queries estimate depth weights to perform a weighted summation of the features at all reference points along the ray, thereby generating the projected image features. Considering that some reference points may exceed the range of scene features, their corresponding weights are set to 0. The process can be formulated as: $\hat{F}^{(v)}_{img} = \sum_{d \in \mathcal{D}_{hit}} W_d \odot \mathrm{DA}(Q_{\mathrm{CFV}}, P_{\mathrm{CFV}}, \hat{F}_{scn})$, where $Q_{\mathrm{CFV}} \in \mathbb{R}^{L_Q \times C}$ are the flattened 3D CFV queries with C = 96, $P_{\mathrm{CFV}} \in \mathbb{R}^{L_Q \times 3}$ denote the corresponding reference points, $\hat{F}_{scn} \in \mathbb{R}^{L_V \times C}$ is the reconstructed scene feature sequence, and the set \mathcal{D}_{hit} refers to the depth positions along the ray where reference points fall within the valid scene feature range, ensuring that only effective depth positions contribute to the aggregated image feature. Here, $L_Q = 60 \times 32 \times 32$ is the CFV query sequence length, and $L_V = \sum_{i=0}^2 (8 \times 128 \times 128 \times 2^{-3i})$ is the total reconstructed scene feature sequence length across resolutions.

Image Decoder progressively restores pixel-level details by processing scene features projected onto the image plane. As its preceding stage, the scene decoder aggregates scene features along the ray depth dimension but lacks interactions between rays. To complement this, it maps the projected scene features (C=96) to 768 dimensions via a linear layer, models global spatial and semantic relationships on the image plane by self-attention, and upscales the resolution from 32×32 to 256×256 with deconvolution, reconstructing fine-grained image details.

3.1.3 Loss

KL Divergence Loss regularizes the latent distribution of the state features, enforcing closeness to a standard normal distribution and ensuring continuity in the latent space: $\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(q_{\phi}(z \mid x) \| p(z)) = \frac{1}{2} \sum_{i=1}^{d} (\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2)$, where p(z) is defined as $\mathcal{N}(0, I)$, d is the dimension of state features, and μ_i , σ_i^2 are the mean and variance of the i-th latent dimension predicted by the encoder E. To allow gradient-based optimization of the stochastic sampling process, the reparameterization trick is used. Instead of directly sampling z from $q_{\phi}(z \mid x)$, it is reparameterized as: $z = \mu + \sigma \odot \epsilon$, $(\mu, \sigma) = E(x)$, $\epsilon \sim \mathcal{N}(0, I)$.

Reconstruction Loss ensures that the reconstructed image $\hat{x} = G(z)$ retains both pixel-level details and high-level semantic structure of the target image x. This is achieved by combining pixel-wise loss with perceptual loss: $\mathcal{L}_{\mathrm{R}} = \mathcal{L}_2 + \mathcal{L}_{\mathrm{perceptual}} = \|x - \hat{x}\|^2 + \sum_l \|\psi_l(x) - \psi_l(\hat{x})\|^2$. Here, \mathcal{L}_2 enforces pixel-wise similarity between the image x and its reconstruction \hat{x} , while $\mathcal{L}_{\mathrm{perceptual}}$ captures structural and semantic consistency by comparing feature maps $\psi_l(x)$ and $\psi_l(\hat{x})$ extracted from the l-th layer of a pre-trained VGG-16. This balance preserves fine details and perceptual coherence, yielding realistic reconstructions.

Discriminator Loss enables the discriminator D to distinguish real images from reconstructed ones, improving its ability to provide meaningful adversarial feedback. With the hinge loss formulation, it is expressed as: $\mathcal{L}_D = \max(0, 1 - D(x)) + \max(0, 1 + D(\hat{x}))$, which encourages the discriminator to assign higher scores to real images and lower scores to reconstructed ones. Hinge loss stabilizes adversarial training by preventing excessively large gradients for confident predictions while ensuring effective feedback for refining reconstruction quality, leading to more stable and efficient optimization.

Adversarial Loss leverages the discriminator's feedback to enhance the perceptual realism of reconstructed images and is defined as: $\mathcal{L}_{A} = -D(\hat{x})$

Total Loss for Encoder and Decoder combines the KL divergence loss, reconstruction loss, and adversarial loss, ensuring effective latent space regularization and perceptual realism. It is formulated as: $\mathcal{L}_G = \beta \cdot \mathcal{L}_{KL} + \mathcal{L}_R + 0.1 \cdot \lambda \cdot \mathcal{L}_A$ where $\beta = 10^{-6}$ controls the strength of the KL divergence regularization. The adaptive weight λ balances the adversarial loss relative to the reconstruction loss, ensuring that the adversarial term contributes meaningfully without overpowering reconstruction. It is computed as $\lambda = \frac{\nabla_{G_L}[\mathcal{L}_R]}{\nabla G_L[\mathcal{L}_A] + \delta}$ with $\nabla G_L[\cdot]$ denoting the gradient of the corresponding term with respect to the last layer L of the decoder, and $\delta = 10^{-6}$ ensuring numerical stability.

271

272273274

275

276

278279

281

284

287

289

291

292

293294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310 311 312

313 314

315 316

317

318

319

320

321

322

323

Figure 3: **Multi-view image reconstruction on nuPlan.** Row 1 shows real images from the nuPlan validation set and Row 2 shows the corresponding reconstructions. Pedestrians, traffic lights, trucks, trailers, cars, crosswalks, and road markings are faithfully reconstructed.

Model	Training	Validation	PSNR ↑	SSIM↑	MVSC↑	rFID↓
SD-VAE	LAION-5B	nuScenes	29.63	0.8283	0.9292	2.18
BEV-VAE BEV-VAE			26.13 28.88	0.7231 0.8028	0.9250 0.9756	6.66 4.74
(a) Reconstruction metrics on nuScenes compared with SD-VAE.						
Model	Training	Validation	PSNR↑	SSIM↑	MVSC↑	rFID↓
SD-VAE	LAION-5B	AV2	27.81	0.8229	0.8962	1.87

Model	Training	Validation	PSNR↑	SSIM↑	MVSC↑	rFID↓
SD-VAE	LAION-5B	AV2	27.81	0.8229	0.8962	1.87
BEV-VAE BEV-VAE	AV2 PAS	AV2 AV2	26.02 27.29	0.7651 0.8028	0.9197 0.9461	4.15 2.82

(b) Reconstruction metrics on AV2 compared with SD-VAE.

Table 2: **BEV-VAE vs. SD-VAE in multi-view reconstruction.** SD-VAE focuses on per-view image fidelity, whereas PAS-trained BEV-VAE achieves superior multi-view spatial consistency (MVSC). 3.2 SPATIALLY-ALIGNED BEV GENERATION FROM 3D OBJECT LAYOUTS

BEV-VAE w/ DiT extends BEV-VAE by integrating DiT in its latent space, leveraging CFG to enhance conditional generation. By explicitly incorporating structured occupancy constraints from 3D object bounding boxes, it ensures spatial consistency and controllability in generation. Given a set of 3D bounding boxes $\{\mathbf{b}_i\}_{i=1}^N$, each parameterized as: $\mathbf{b} = (q_w, q_x, q_y, q_z, x_c, y_c, z_c, l, w, h, c)$, where the quaternion $q=(q_w,q_x,q_y,q_z)$ encodes the 3D orientation, (x_c,y_c,z_c) specifies the box center in the ego coordinate system, (\bar{l}, w, h) represents the size of the box, and $c \in 1, \ldots, C$ is the semantic class index. These boxes are voxelized into a binary occupancy tensor $\mathbf{C}_{box} \in \{0,1\}^{C \times 8 \times 128 \times 128}$, where each voxel represents whether a given spatial location is occupied by a bounding box of a particular class. Formally, it is defined as: $C_{\text{box}}(c, z, y, x) = \max_{i:c_i=c} \mathbf{1}[(z, y, x) \in \Omega(\mathbf{b}_i)]$ where $1|\cdot|$ is an indicator function, and $\Omega(\mathbf{b}_i)$ denotes the discretized voxelized representation of bounding box b_i . The max operation aggregates occupancy information from overlapping bounding boxes within the same class. The occupancy tensor $C_{\rm box}$ is downsampled via non-overlapping patch partitioning in the BEV plane, yielding a feature of shape $96 \times 8 \times 32 \times 32$, followed by channel-wise concatenation of the height dimension to form the conditional occupancy feature $F_{box} \in \mathbb{R}^{768 \times 32 \times 32}$. Aligned with the state feature $F_{\text{stt.}}$, it is injected via element-wise addition: $F'_{\text{stt.}} = F_{\text{stt.}} + s \cdot F_{\text{loox}}$, where s is the guidance scale in CFG. This ensures spatial consistency by aligning the conditional occupancy features and state features within the shared BEV coordinate system, allowing DiT to focus on relevant regions by explicitly incorporating object category and location information.

4 EXPERIMENTS

4.1 DATASETS

This study uses four multi-camera autonomous driving datasets that differ substantially in scale, camera configuration, annotated categories, and recording locations, as shown in Tab. 1. Despite these differences, all datasets provide full 360° coverage of the surrounding scene.

The WS101 dataset (Zürn et al., 2024) consists of 5 cameras with 101 scenes. We use the first 84 scenes as the training set and the remaining 17 scenes as the validation set. Each scene contains approximately 200 samples. Note that 3D object bounding boxes are not provided.

The nuScenes dataset (Caesar et al., 2020) consists of 6 cameras with 700 training scenes and 150 validation scenes. Each scene contains approximately 220 samples, of which 40 are annotated across

Figure 4: **Novel view synthesis via camera pose modifications on nuScenes.** Row 1 shows real images from the nuScenes validation set, and Rows 2-3 show reconstructions with all cameras rotated 30° left and right, where the cement truck and tower crane truck remain consistent across views without deformation.



Figure 5: **Novel view synthesis cross camera configurations.** Row 1 presents real images from the nuPlan validation set. Row 2 and Row 3 show reconstructions using camera parameters from AV2 and nuScenes, respectively. The model captures dataset-specific vehicle priors: AV2 reconstructions include both the front and rear of the ego vehicle, while nuScenes reconstructions mainly show the rear (with the rightmost image corresponding to the rear-view camera for alignment).

10 object categories. In total, it includes 155k training samples, of which 28k are annotated, and 33k validation samples, of which 6k are annotated.

The AV2 dataset (Wilson et al., 2023) consists of 7 cameras, with the front camera rotated by 90°. It includes 700 training scenes and 150 validation scenes. Each scene contains approximately 300 samples, of which 150 are annotated across 30 object categories. In total, it includes 224k training samples, of which 109k are annotated, and 47k validation samples, of which 23k are annotated.

The nuPlan dataset (Caesar et al., 2021) consists of 8 cameras with 1085 training logs. The training set comprises 3.11 million samples annotated with 7 object categories, but we only use the images from the training set.

4.2 Settings

We introduce a new hybrid autonomous driving dataset configuration, **PAS**, which combines nu**P**lan, **AV**2, and nu**S**cenes. The training process consists of two stages, all using the AdamW optimizer with a learning rate of 1e-4 and a 5k-step warm-up.

Stage 1: Training is performed on **PAS** with a batch size of 1 per GPU for 800k iterations on 8 NVIDIA H100 GPUs. The optimization settings are $\beta = (0.9, 0.99)$, weight decay 1e-4, and EMA decay 0.9999.

Stage 2: Training is conducted on **PAS** with a batch size of 8 per GPU for 200k iterations with 3D object annotations from AV2 or 400k iterations with annotations from nuScenes, using 8 NVIDIA A800 GPUs. The optimization settings are $\beta = (0.9, 0.95)$, weight decay 0.1, and EMA decay 0.999.

4.3 METRICS

The performance of BEV-VAE is evaluated using multiple metrics covering reconstruction quality, multi-view spatial consistency, and generation quality.

PSNR and **SSIM** measure the similarity between reconstructed and original images, with PSNR assessing signal fidelity and SSIM focusing on structural consistency.

Ground Truth (WS101 cams)

Reconstructed (Zero-shot)

Reconstructed (Fine-tune)

Figure 6: **Zero-shot BEV representation construction on WS101.** Row 1 shows real images from the WS101 validation set. Rows 2 and 3 show zero-shot and fine-tuned reconstructions, respectively, with object shapes preserved in the zero-shot results and further sharpened after fine-tuning.

Model	Training	Validation	PSNR↑	SSIM↑	MVSC↑	rFID↓
SD-VAE	LAION-5B	WS101	23.38	0.7050	0.8580	4.59
BEV-VAE BEV-VAE	PAS PAS+WS101	WS101 WS101	16.6 23.46	0.3998 0.6844	0.8309 0.9505	56.7 13.78

Table 3: **Zero-shot and fine-tuned reconstruction metrics on WS101 compared with SD-VAE. Multi-View Spatial Consistency (MVSC)** evaluates spatial consistency in multi-view reconstruction. Following BEVGen Swerdlow et al. (2024) and DriveWM Wang et al. (2024), a pre-trained LoFTR Sun et al. (2021) is used to compute keypoint matching confidence between adjacent views. MVSC is the ratio of average adjacent-view matching confidence in reconstructed images to that in real images, where higher values imply better alignment.

FID and **FVD** are used to evaluate the quality of generated data in a deep feature space. FID measures the fidelity of reconstructed and generated multi-view images, while FVD assesses the temporal consistency and realism of generated front-view video sequences.

4.4 MULTI-VIEW IMAGE RECONSTRUCTION

BEV-VAE learns unified BEV representations by reconstructing multi-view images, integrating semantics from all camera views while modeling 3D spatial structure. Reconstruction metrics provide an indirect evaluation of the quality of the learned BEV representations. For reference, we compare with SD-VAE (Rombach et al., 2022), a foundational model trained on LAION-5B Schuhmann et al. (2022), which encodes a single 256×256 image into a $32 \times 32 \times 4$ latent. In contrast, BEV-VAE encodes multiple 256×256 views into a $32 \times 32 \times 16$ BEV latent, facing the more challenging task of modeling underlying 3D structure. As shown in Tab. 2, BEV-VAE trained on nuScenes or AV2 alone underperforms SD-VAE. However, when trained on the hybrid PAS dataset that combines multiple autonomous driving datasets with diverse camera configurations, BEV-VAE achieves a notable improvement, surpassing SD-VAE by a large margin on MVSC. This demonstrates that BEV-VAE effectively integrates multi-view semantics and captures spatial structure. Moreover, as illustrated in Fig. 3, BEV-VAE reconstructs most elements of complex driving scenes with high fidelity, while decoupling per-view reconstruction quality from cross-view spatial consistency: since all views are decoded from the same BEV representation, spatial consistency across views is guaranteed regardless of per-view reconstruction quality.

4.5 NOVEL VIEW SYNTHESIS

Reconstruction metrics provide a quantitative proxy for evaluating the quality of BEV representations, but they cannot directly verify whether BEV-VAE accurately models the spatial structure of objects and background from multi-view semantics. Conversely, if BEV-VAE captures such spatial structures correctly, it should be able to synthesize novel views simply by adjusting the camera poses, as illustrated in Fig. 4. Furthermore, leveraging the hybrid PAS dataset configuration, BEV-VAE demonstrates generalization not only to unseen camera poses but also to varying camera intrinsics, enabling the reconstruction of nuPlan scenes under the camera configurations of AV2 or nuScenes, as shown in Fig. 5.

Metric	BEVGen	Panacea	MagicDrive	DrivingDiffusion	DriveWM	Ours
gFID↓	25.54	16.96	16.20	15.83	12.99	20.7
MVSC↑	-	0.9189	0.8310	-	-	0.9310

Table 4: Comparison of multi-view image generation on nuScenes.

Dataset	0	1	2	3	4	5	6	7	8	9
nuScenes AV2					20.70 16.06					

Table 5: Impact of CFG scale on gFID for multi-view image generation.

4.6 ZERO-SHOT BEV REPRESENTATION CONSTRUCTION

Given that BEV-VAE already demonstrates generalization across camera configurations, it further exhibits a degree of zero-shot BEV representation construction capability when applied to multi-view images with camera setups unseen during training. We validate this property on WS101, as illustrated in Fig. 6. Moreover, benefiting from the strong spatial priors learned by BEV-VAE, fine-tuning on WS101 with a new camera configuration for 50k iterations leads to a significant improvement in reconstruction metrics, surpassing SD-VAE in both PSNR and MVSC, as shown in Tab. 3.

4.7 AUTONOMOUS DRIVING SCENE SYNTHESIS

As shown in Fig. 1, BEV-VAE w/ DiT generates BEV representations from 3D object layouts that can be decoded to arbitrary viewpoints, enabling a single model to support vehicles with different camera setups and achieve cross-platform scene generalization. We compare our approach with prior multi-view image generation methods in Tab. 4. Although our method has a higher gFID than previous works, it demonstrates superior multi-view spatial consistency. CFG scale ablation (Tab. 5) shows that the optimal gFID is achieved at a scale of 4 for nuScenes (20.7) and 6 for AV2 (15.73).

4.8 Data Augmentation for Perception

BEV-VAE w/ DiT using the Historical Frame Replacement strategy (randomly replacing real frames with generated ones) improves BEVFormer's perception by enabling the model to learn invariance of object locations relative to appearance. Compared to BEVGen, which augments the dataset by adding synthetic data, our approach requires no additional computational cost while achieving the highest NDS, as shown in Tab. 6.

Perception Model	Generative Model	Augmentation Strategy	mAP↑	NDS↑
BEVFormer Tiny	-	-	25.2	35.4
BEVFormer Tiny	BEVGen	Training Set + 6k Synthetic Data	27.3	37.2
BEVFormer Tiny	BEV-VAE w/ DiT	Historical Frame Replacement	27.1	37.4

Table 6: Perception performance with generative augmentation.

5 CONCLUSION

In this paper, we present BEV-VAE, a variational autoencoder that unifies multi-view images into a compact BEV representation, capturing both scene semantics and 3D structure. BEV-VAE supports encoding from arbitrary camera layouts and decoding to any desired viewpoints, enabling scalable training across datasets with different camera configurations. Within the latent space of BEV-VAE, DiT can generate BEV representations conditioned on 3D object layouts, which can also be decoded to arbitrary viewpoints, allowing cross-platform generalizable applications. Moreover, this synthesized imagery significantly enhances the performance of downstream perception models. Although BEV-VAE does not surpass previous methods in FID for multi-view image synthesis, this is partly due to the greater difficulty of generating full scenes compared with fixed-view images. Another limitation is the relatively low resolution (256×256), which we believe can be addressed via super-resolution models, allowing spatial modeling and resolution enhancement to be handled separately. In the future, we plan to extend BEV-VAE to temporal scenarios.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- 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*, pp. 11621–11631, 2020.
- Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv* preprint arXiv:2106.11810, 2021.
- Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv* preprint arXiv:2410.10733, 2024.
- Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.
- 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.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- 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*, pp. 17853–17862, 2023.
- Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.
- Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pp. 469–485. Springer, 2024a.
- Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.
- Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.
- 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), pp. 2774–2781. IEEE, 2023.
- 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*, 65(1):99–106, 2021.

- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.
 - 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*, pp. 194–210. Springer, 2020.
 - 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*, pp. 10684–10695, 2022.
 - Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.
 - Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv* preprint arXiv:2010.02502, 2020.
 - Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8922–8931, 2021.
 - Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 2024.
 - Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
 - 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024.
 - Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024.
 - Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.
 - Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.
 - Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.
 - 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*, pp. 3836–3847, 2023.
 - Jannik Zürn, Paul Gladkov, Sofia Dudas, Fergal Cotter, Sofi Toteva, Jamie Shotton, Vasiliki Simaiaki, and Nikhil Mohan. Wayvescenes101: A dataset and benchmark for novel view synthesis in autonomous driving. *arXiv preprint arXiv:2407.08280*, 2024.

Supplementary Material for BEV-VAE

The supplementary material offers additional context and results that enhance the main paper on BEV-VAE. First, Sec. A provides the core principles of the generative models used in our framework. Then, Sec. B then explains the multi-view spatial consistency (MVSC) metric in detail and compares it with prior methods. In Lastly, Sec. C presents examples of fine-grained 3D object layout control, enabling adjustments in the number and position of vehicles.

A Preliminary for Generative Models

VAE is trained by maximizing the Evidence Lower Bound (ELBO) as follows:

$$\log p_{\theta}(x) \ge \mathbb{E}_{q_{\phi}(z|x)} \left[\log p_{\theta}(x|z) \right] - D_{\mathrm{KL}} \left(q_{\phi}(z|x) \parallel p_{\theta}(z) \right), \tag{1}$$

where x is the input data, z is the latent variable, ϕ and θ are the encoder and decoder parameters, respectively. The first term ensures that the decoder $p_{\theta}(x \mid z)$ can accurately reconstruct x from the latent variable z, and the second term penalizes the divergence between the posterior $q_{\phi}(z \mid x)$ and the prior p(z), typically $\mathcal{N}(0, I)$, encouraging a structured and continuous latent space.

Diffusion models define a forward process that gradually adds Gaussian noise to real data x_0 , formulated as:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t$ are pre-defined noise scheduling coefficients, enabling direct sampling of x_t from x_0 without iterative noise application. With reparameterization, the noised sample is:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{I}).$$
 (3)

This highlights the relationship between x_0 and noise ϵ_t , enabling training via noise prediction. The reverse process learns to iteratively denoise x_t back to x_0 , where

$$p_{\theta}(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t), \sigma_t^2 \mathbf{I}), \tag{4}$$

The mean $\mu_{\theta}(x_t)$ is predicted by the model, while the variance σ_t^2 is fixed as in DDPM. The ELBO is minimized during training, simplifying to a noise prediction objective:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}[\|\epsilon_{\theta}(x_t) - \epsilon_t\|_2^2]. \tag{5}$$

Sampling starts from a standard Gaussian $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises via $p_{\theta}(x_{t-1} \mid x_t)$ to generate samples consistent with the target distribution.

Classifier-Free Guidance (CFG) enhances conditional diffusion models by adjusting the sampling process to prioritize samples with high $p(c \mid x)$. By applying Bayes' rule, the gradient formulation is derived as:

$$\nabla_x \log p(c \mid x) = \nabla_x \log p(x \mid c) - \nabla_x \log p(x), \tag{6}$$

which implies that increasing $p(c \mid x)$ can be achieved by adjusting the diffusion trajectory toward higher $p(x \mid c)$. The reverse diffusion process follows:

$$p_{\theta}(x_{t-1} \mid x_t, c) = \mathcal{N}(x_{t-1} \mid \mu_{\theta}(x_t, c), \sigma_t^2 \mathbf{I}).$$
 (7)

To guide the diffusion towards the conditional distribution, CFG modifies the noise prediction as:

$$\hat{\epsilon}_{\theta}(x_t, c) = \epsilon_{\theta}(x_t, \emptyset) + s \cdot (\epsilon_{\theta}(x_t, c) - \epsilon_{\theta}(x_t, \emptyset)) \propto \epsilon_{\theta}(x_t, \emptyset) + s \cdot \nabla_x \log p(c \mid x_t). \tag{8}$$

During training, conditioning is randomly dropped to learn both conditional and unconditional noise predictions.

B EVALUATION WITH MULTI-VIEW SPATIAL CONSISTENCY

Evaluating images with pre-trained models is a common practice, with metrics such as Inception Score (IS), Fréchet Inception Distance (FID), and Learned Perceptual Image Patch Similarity (LPIPS) widely used. To assess spatial consistency in multi-view generation, a matching-based metric is introduced. Following prior works such as BEVGen and DriveWM, a pre-trained LoFTR model is employed to perform keypoint matching between adjacent views. Given that the overlapping regions

Method	FID↓	MVSC↑	Object Layouts	Camera Poses	Other Conditions
MagicDrive Panacea	16.20 16.96	0.8310 0.9189	Fourier embedding(1D) Perspective projection (2D)	Fourier embedding Pseudo-color image	Text, map. Text, map, depth.
Ours	20.70	0.9310	Binary occupancy (3D)	Extrinsic matrix	None

Table 7: Comparison on nuScenes: image quality, spatial consistency, and conditions



Figure 7: **Multi-View Spatial Consistency (MVSC) on nuScenes.** The comparison is based on images generated by different methods. Row 1 shows the projections of 3D object layouts onto the image plane. Row 2 presents the corresponding validation images. Rows 3–5 display the results generated by MagicDrive, Panacea, and our method, respectively. To better visualize spatial consistency across adjacent views, each row of images is shifted to the right by half an image width. Vertical black lines mark the centerlines of each camera view. Red boxes indicate regions where the generated vehicles are significantly misaligned with the ground-truth layouts.

between adjacent views typically cover no more than half of the image centered horizontally, each image is divided vertically into left and right halves. For each adjacent camera pair, keypoint matching is performed between the two bordering half-images, as shown in Fig. 7. The proposed Multi-View Spatial Confidence (MVSC) is then defined as the ratio of this average confidence from reconstructed or generated images to that from real images, serving as an indicator of spatial consistency across views.

Based on the same MVSC metric, Table 7 compares MagicDrive, Panacea, and our method. While our approach yields a higher FID on nuScenes than prior methods, it achieves the best spatial consistency. BEV-VAE adopts a more direct and physically grounded representation of object layouts. MagicDrive encodes 3D boxes with Fourier embeddings and MLPs, fusing them with image features via cross-attention. Panacea projects 3D boxes into the image plane and enforces pixel-level alignment using ControlNet. In contrast, our method represents object layouts as binary occupancy maps in the BEV space, which are inherently aligned with the 3D BEV representation without requiring additional projection or alignment. Camera poses are also handled in a physically consistent way: by rotating the extrinsic matrix applied to the BEV representation, novel views can be rendered directly. This principled 3D-to-2D mapping preserves spatial relationships across views, leading to inherently consistent multi-view generation.

C GENERATION WITH PRECISE 3D OBJECT CONTROL

To demonstrate that the BEV latent space supports precise control based on structured 3D object layouts, we generate multi-view images by selectively removing different vehicles from the same scene. As shown in Fig. 8, Row 1 presents real images from the validation set, and Row 2 shows the reconstructed images. Row 3 displays images generated from the corresponding 3D bounding boxes. Rows 4–8 further illustrate controllable generation by selectively removing specific vehicles from the input layouts, with the removed objects indicated by numerical labels.

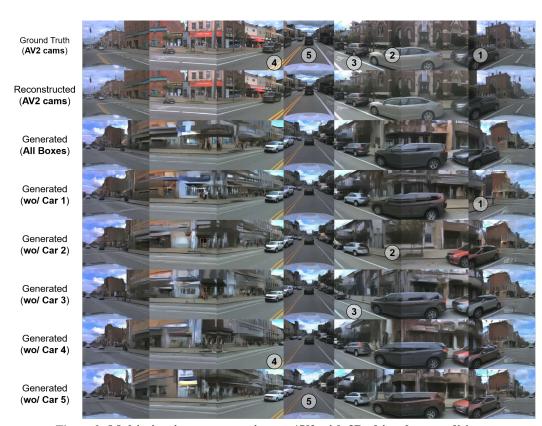


Figure 8: Multi-view image generation on AV2 with 3D object layout editing.