# ANYSCENE: TOWARDS HIGHLY CONTROLLABLE DRIVING SCENE GENERATION AT ANYWHERE AND BEYOND

**Anonymous authors**
Paper under double-blind review




Figure 1: **Scalable and controllable driving scene generation with AnyScene.** AnyScene generates semantic occupancy zero-shot from user-designed BEV layouts and layouts derived from unseen datasets and global OpenStreetMap data (OpenStreetMap contributors, 2025). The generated occupancy grounds novel-view video synthesis under flexible camera rigs, and the added views densify sparse-view 3D reconstruction.

## ABSTRACT

Recorded driving datasets cannot exhaustively cover the various road layouts, traffic flows, and rare safety-critical scenarios needed for end-to-end autonomous driving. Scaling synthetic data beyond recorded scenes therefore requires a generator that can faithfully realize user-specified layouts, maintain temporally coherent geometry, and render scenes from flexible viewpoints without matching real-world footage. Existing methods struggle to meet these requirements together: end-to-end video generators offer limited geometric control, while occupancy-guided approaches may deviate from BEV layouts and remain constrained by low-frequency signals, reference frames, or fixed camera rigs. We propose **AnyScene**, a unified occupancy-centric framework that autoregressively generates semantic occupancy sequences from diverse BEV layouts and uses the resulting geometry to synthesize temporally consistent multi-view driving videos. AnyScene follows user-designed layouts and generalizes zero-shot to layouts from unseen datasets and global maps, while supporting reference-free video synthesis and flexible camera configurations. To support the framework, we build **nuCraftv2**, a 12 Hz dataset based on nuScenes with synchronized BEV layouts and high-quality dense semantic occupancy. Experiments demonstrate zero-shot occupancy generation across these layout sources and state-of-the-art video generation, together with applications to counterfactual observation generation and sparse-view 3D reconstruction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

# 1 INTRODUCTION

Real-world driving data collection cannot exhaustively capture the diversity of road layouts, traffic flows, and rare safety-critical scenarios needed for end-to-end autonomous driving (Caesar et al., 2020; Sun et al., 2020; Wilson et al., 2021). Scalable and controllable driving data generation can extend coverage beyond recorded scenes, but it requires more than producing realistic videos. A generator must faithfully realize layouts and traffic flows specified through diverse inputs, maintain coherent geometry at the temporal resolution needed for video, and render the resulting scenes from flexible viewpoints without matching real-world footage. Meeting these requirements together remains an open challenge.

Existing approaches satisfy only parts of this goal. End-to-end video generation models (Hu et al., 2023a; Russell et al., 2025; Gao et al., 2024b; 2025; Zhao et al., 2024; Wen et al., 2024) produce visually compelling sequences, but their implicit scene representations make fine-grained control and geometric consistency across views and time difficult. Reconstruction-based simulators (Wu et al., 2023; Yang et al., 2023a; Tonderski et al., 2024; Yan et al., 2024; Zhou et al., 2024; Chen et al., 2025b) recover detailed geometry from recorded observations, but their dependence on per-scene data and optimization limits generation beyond captured environments. Generating an explicit scene representation from a portable control interface offers a way to specify new geometry independently of recorded imagery.

3D semantic occupancy provides a structured, editable, and view-independent representation of scene geometry and semantics (Cao & de Charette, 2022; Huang et al., 2023b; Zheng et al., 2023; Wang et al., 2024). BEV layouts provide the complementary control interface: roads and agent configurations can be expressed in a shared format derived from different datasets, maps, or user designs (Hu et al., 2023b; Gao et al., 2024a; Li et al., 2025). Together, they allow users to specify new scenes and counterfactual traffic flows without corresponding real-world recordings. Recent occupancy-guided frameworks such as UniScene (Li et al., 2025) and InfiniCube (Lu et al., 2025) explore this direction, but still fall short of the requirements above. BEV-to-occupancy conditioning may fail to preserve specified layouts and agents (Yang et al., 2025d; Lee et al., 2024), weakening control over the generated scene. Low-frequency (2 Hz) (Caesar et al., 2020; Yang et al., 2025d; Tian et al., 2023) or coarse conditioning signals (Li et al., 2025) limit precise frame-level control for video generation. Dependence on reference frames restricts generation from layouts without matching imagery, while fixed camera rigs limit viewpoint flexibility (Li et al., 2025; Lu et al., 2025; Yang et al., 2025d).

We propose **AnyScene**, a unified occupancy-centric framework for scalable and controllable driving data generation. From BEV layouts spanning user designs, unseen datasets, and global OpenStreetMap data (OpenStreetMap contributors, 2025), it autoregressively generates semantic occupancy sequences in a compact BEV-based latent space, enabling precise, high-frequency control and zero-shot generalization across diverse layout sources. Using the generated occupancy as a shared geometric representation, AnyScene synthesizes temporally and cross-view consistent multi-view videos without real reference frames and with flexible camera counts and poses. To support the framework, we build *nuCraftv2*, a 12 Hz dataset based on nuScenes with synchronized BEV layouts and high-quality dense semantic occupancy. Experiments demonstrate zero-shot occupancy generation across these layout sources, while the video generator achieves state-of-the-art performance. We further generate counterfactual driving scenarios from user-specified BEV layout sequences specifying alternative traffic flows, while also supporting occupancy-level edits and ego-trajectory perturbations. The generated views augment sparse camera rigs and improve 3D reconstruction completeness.

In summary, the main contributions of this work are as follows:

- We propose AnyScene, a framework that generates controllable semantic occupancy sequences and temporally consistent multi-view videos from diverse BEV layouts. Its occupancy generator generalizes zero-shot across user-designed layouts, unseen datasets, and global maps, while its video generator supports reference-free synthesis with flexible camera configurations and achieves state-of-the-art performance.
- We build *nuCraftv2*, a 12 Hz dataset with synchronized BEV layouts and high-quality dense semantic occupancy, providing high-frequency geometric supervision for the framework.

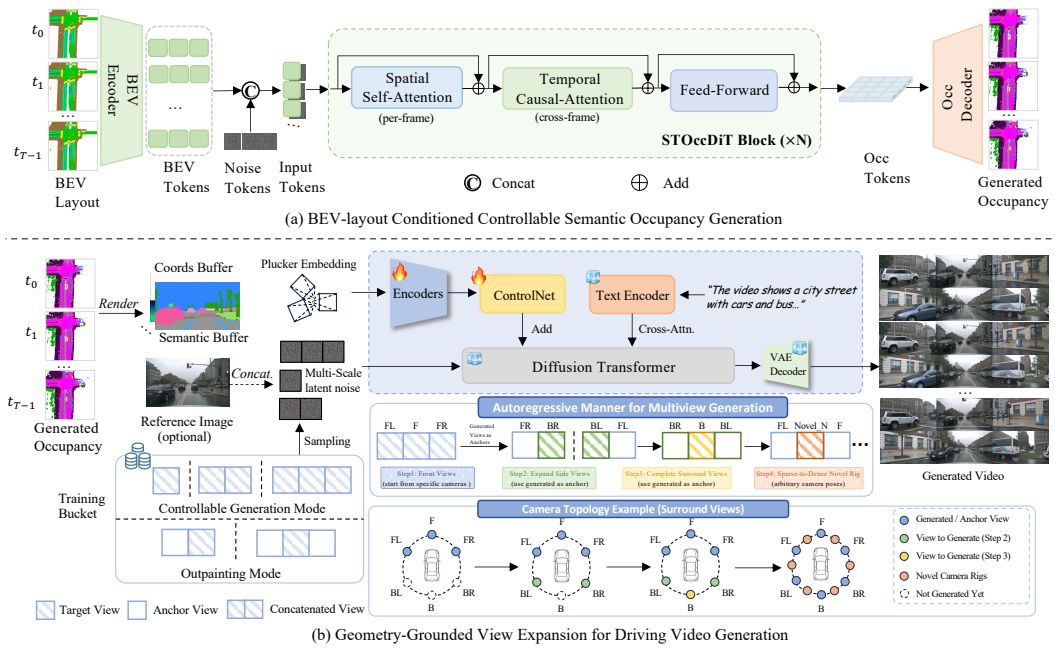


Figure 2: **Overview of AnyScene. (a) BEV-conditioned occupancy generator:** a spatio-temporal diffusion model maps BEV layout sequences to temporally coherent semantic occupancy. **(b) Geometry-grounded view-expansion module:** the generated occupancy provides semantic and coordinate buffers that, together with camera-pose embeddings, condition video synthesis. Autoregressive view expansion produces temporally and cross-view consistent videos for flexible camera counts and poses, with optional real reference frames.

- We explore two applications of AnyScene: generating counterfactual driving scenarios from user-specified BEV sequences that encode novel trajectories and scene edits, and improving sparse-view 3D reconstruction with generated views.

## 2 Method

### 2.1 Problem Formulation

Given a sequence of BEV semantic layouts $\mathcal{B}_{1:T}$, query camera poses $\mathbf{C}_{1:T,1:K}$, a text condition $\mathbf{y}$, and an optional reference image $\mathbf{R}$, our goal is to generate a temporally consistent 3D semantic occupancy sequence $\mathcal{O}_{1:T}$ and the corresponding multi-view driving video sequence $\mathbf{V}_{1:T,1:K}$. We factorize their joint distribution into two stages with semantic occupancy as the central representation:

$$p(\mathcal{O}_{1:T}, \mathbf{V}_{1:T,1:K} \mid \mathcal{B}_{1:T}, \mathbf{C}_{1:T,1:K}, \mathbf{y}, \mathbf{R}) = p_\theta(\mathcal{O}_{1:T} \mid \mathcal{B}_{1:T}) p_\phi(\mathbf{V}_{1:T,1:K} \mid \mathcal{O}_{1:T}, \mathbf{C}_{1:T,1:K}, \mathbf{y}, \mathbf{R}). \tag{1}$$

The optional reference image $\mathbf{R}$ is omitted in the reference-free setting.

In the first stage, a BEV-conditioned occupancy generator produces the occupancy sequence from the input layouts. It supports precise control from user-designed, cross-dataset, and global-map BEV inputs while maintaining temporal consistency. In the second stage, we synthesize multi-view driving videos directly from the generated occupancy sequence. A geometry-grounded view-expansion module takes $\mathcal{O}_{1:T}$ as input and produces a temporally and cross-view consistent video sequence $\mathbf{V}_{1:T,1:K}$, where both the number of camera views $K$ and the camera poses can be flexibly specified at inference time.

This occupancy-centric formulation uses BEV layouts as the sole scene-structure input, while camera poses specify the queried views and text or an optional reference image controls appearance.

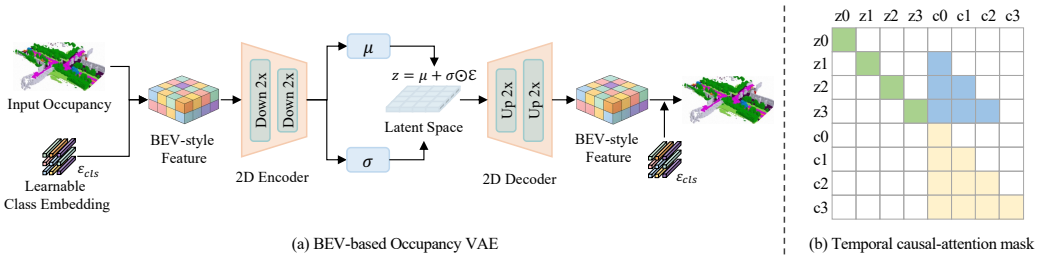


Figure 3: **Occupancy representation and temporal modeling.** (a) The BEV-based occupancy VAE compresses 3D semantic occupancy into a compact 2D latent while preserving height-dependent semantics. (b) The dual-stream causal-attention mask exposes clean past latents to the current noisy latent during parallel training; generated latents serve as history during autoregressive inference.

### 2.2 Controllable Semantic Occupancy Generation

We encode each 3D semantic occupancy frame into a compact 2D latent with a BEV-based occupancy VAE, whose 2D encoder and decoder preserve height-dependent semantic information while providing an efficient latent grid for sequence modeling. A multi-hot BEV encoder embeds the active layout classes and spatially aligns the resulting tokens with the occupancy latents. Instead of injecting the BEV layout as an external condition, our occupancy generator jointly processes the aligned BEV and noisy occupancy tokens through spatial attention, allowing layout constraints to interact directly with the evolving occupancy representation. It then applies causal temporal attention under a dual-stream mask, which exposes clean past latents during parallel training and replaces them with previously generated latents during autoregressive inference. This design combines precise layout adherence with stable temporal propagation for high-frequency occupancy generation. We train the latent generator with rectified flow matching and decode its outputs into semantic occupancy sequences. Figure 3 illustrates the VAE and dual-stream causal mask; further architectural and training details are provided in Appendix C.2–C.3.

### 2.3 Geometry-Grounded View Expansion

Existing driving video diffusion models take text, single images, reference clips, or abstract geometric controls such as 3D bounding boxes and BEV layouts directly as generation conditions. These inputs can specify appearance or layout, but do not explicitly represent dense 3D scene geometry for consistent synthesis across views and flexible camera configurations. To overcome this limitation, we use the geometry-grounded view-expansion module to generate temporally and cross-view consistent multi-view driving videos from occupancy sequences, as depicted in Fig. 2 (b).

The view-expansion module treats the generated volumetric occupancy as the canonical 3D scene representation. For each camera view, three geometric signals are rendered from the occupancy volume using $f$VDB (Williams et al., 2024), following InfiniCube (Lu et al., 2025): a semantic buffer, a coordinate buffer, and the Plücker embedding (Yang et al., 2026) of the camera pose. These explicit geometric cues enable the model to perform geometry-driven video synthesis, where the camera rig becomes a flexible, query-time variable.

The view-expansion module builds upon the Wan2.1-T2V-14B (Wan et al., 2025) backbone and adopts a VACE-style (Jiang et al., 2025b) ControlNet architecture, where a trainable control branch runs alongside the frozen DiT backbone to process per-camera geometric conditions, including reference-view RGB latents, Plücker camera embeddings, and semantic/coordinate buffer features. The resulting hint tensors are propagated back into the backbone via uniformly spaced residual connections, preserving the pretrained generative prior while enabling precise geometric control over generation. Text prompts describing the driving environment are encoded as additional conditioning inputs to control scene attributes, and an optional reference camera image can further be provided to enforce appearance consistency. The model is trained with a unified mode-mixture strategy that jointly optimizes multi-view video generation with or without a reference view, as well as various outpainting configurations, while keeping the DiT backbone, text encoder, and VAE frozen throughout. At inference, the module synthesizes surround-view videos *autoregressively* from the

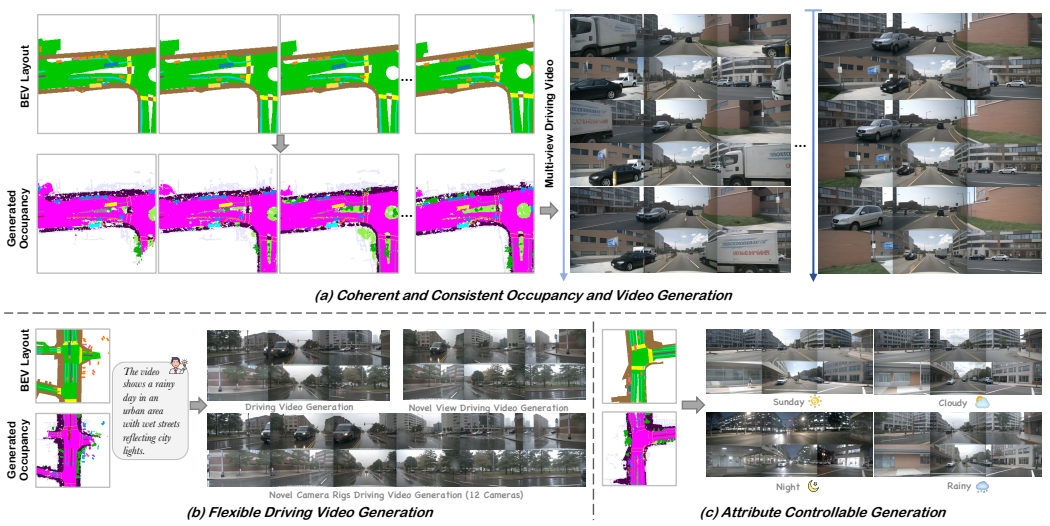


Figure 4: **Versatile generation ability of AnyScene.** (a) Coherent and temporally consistent occupancy and multi-view video generation conditioned on BEV layout sequences. (b) Flexible driving video generation from user prompts and occupancy, including novel views and flexible camera rigs. (c) Controllable generation of attribute-diverse videos by modifying the input text prompts. Please zoom in for better results.

occupancy sequence alone: starting from the front cameras, the model progressively expands to the remaining views by leveraging previously generated frames as anchors, a mechanism that supports flexible camera rigs by allowing virtual views to be inserted at desired poses. Grounding video generation in explicit 3D occupancy enables flexible, reference-free synthesis of multi-view driving videos across camera counts and poses within a single unified model.

# 3 Experiments

## 3.1 Datasets

**nuCraftv2.** To support AnyScene, we build *nuCraftv2*, a high-quality, high-resolution, high-frequency dataset derived from nuScenes (Caesar et al., 2020). Its key design is to reconstruct the static scene from aggregated LiDAR sweeps and model dynamic agents with per-instance 3D assets, then assemble both using per-frame object poses and HD-map priors. This yields dense, instance-aware semantic occupancy at voxel sizes as fine as 0.1 m, synchronized with BEV layouts and multi-view camera images at 12 Hz. The curation pipeline and experimental grid configuration are detailed in Appendix B.

## 3.2 Implementation Details

The BEV-conditioned occupancy generator is trained using PyTorch based on the mmengine (Contributors, 2022) framework. The effective batch size is set to 2 per GPU for the occupancy generator and 8 single-frame grids per GPU for the BEV-based occupancy VAE. We train the geometry-grounded view-expansion module for 5k iterations at resolution $480 \times 832$ using 49 sampled frames per iter. Optimization is performed with AdamW using a learning rate of $2 \times 10^{-5}$, together with bf16 mixed precision and gradient checkpointing. Training follows a curriculum over single-view, two-view, and three-view T2V settings, as well as outpainting modes. More implementation details of the BEV-based occupancy VAE, occupancy generator, and view-expansion module are provided in the supplementary material. We use eight NVIDIA A800 GPUs with 80 GB memory each for occupancy generation and NVIDIA H200 GPUs for video generation.

We evaluate AnyScene in four aspects. **1) Occupancy Generation:** Following (Yang et al., 2025c), we report mIoU and IoU for VAE reconstruction and occupancy generation in 3D, BEV, and BEV-

Table 1: **Occupancy VAE reconstruction.** Downsampled size denotes the latent shape $(C \times H \times W)$. Best results are in **bold**.

| Method | Downsampled Size | mIoU ↑ | IoU ↑ |
|---|---|---|---|
| OccLLama (VQVAE) | $128 \times 50 \times 50$ | 75.2 | 63.8 |
| OccWorld (VQVAE) | $128 \times 50 \times 50$ | 66.4 | 62.3 |
| OccSora (VQVAE) | $512 \times 25 \times 25$ | 27.4 | 37.0 |
| $\mathcal{X}$-Scene (Triplane-VAE) | $32 \times 64 \times 64$ | 58.05 | 44.96 |
| $\mathcal{X}$-Scene (Triplane-VAE) | $32 \times 128 \times 128$ | 81.47 | 61.94 |
| UniScene (VAE) | $8 \times 50 \times 50$ | 72.90 | 64.10 |
| **Ours** (VAE) | $32 \times 64 \times 64$ | 76.02 | 62.30 |
| **Ours** (VAE) | $16 \times 128 \times 128$ | 85.29 | 73.72 |
| **Ours** (VAE) | $32 \times 128 \times 128$ | **90.27** | **81.25** |

Table 2: **Comparison of video generation methods** on FVD, mIoU, and mAP.

| Method | Frames | Cond. | FVD ↓ | mIoU ↑ | mAP ↑ |
|---|---|---|---|---|---|
| Panacea | 8 | BEV | 139.00 | - | - |
| Vista | 25 | Action | 112.65 | - | - |
| OmniNWM | 16 | Action | 60.49 | 34.98 | 22.92 |
| MagicDrive | 60 | BEV & 3D Box | 217.94 | 18.27 | 11.86 |
| MagicDrive-V2 | 16 | BEV & 3D Box | 94.84 | 20.40 | 18.17 |
| WoVoGen | 6 | Occ | 417.70 | - | - |
| UniScene | 8 | Occ | 70.52 | 21.75 | 10.32 |
| UniScene-Rollout | 32 | Occ | 610.15 | 18.69 | 6.24 |
| GenieDrive | 8 | Occ | 55.93 | 30.97 | 19.17 |
| | 16 | | 63.65 | | |
| | 49 | | 92.52 | | |
| **Ours** | 8 | Occ | **35.98** | **38.26** | **27.73** |
| | 16 | | **43.19** | | |
| | 49 | | **57.57** | | |

to-layout settings, together with FID and KID on camera- and BEV-view renderings. **2) Video Generation:** We measure video quality with FVD and controllability with BEV segmentation and 3D object detection on generated nuScenes (Caesar et al., 2020) validation videos. **3) Counterfactual Scenario Generation:** We qualitatively assess arbitrary traffic flows, occupancy edits, and ego-trajectory changes for edit adherence and preservation of unedited content. **4) Sparse-View 3D Reconstruction:** We augment sparse real cameras with generated novel views and evaluate two reconstruction models using LiDAR-referenced accuracy, completeness, recall, F-score, and BEV coverage. Further metric details are provided in the supplementary material.

### 3.3 Main Results

**Versatile Generation Ability.** As shown in Fig. 4, AnyScene supports versatile and controllable driving scene generation. Given a BEV layout sequence, the BEV-conditioned occupancy generator produces a temporally coherent semantic occupancy sequence, which then conditions the geometry-grounded view-expansion module to produce multi-view videos. The resulting occupancy and videos share the same geometric structure, ensuring consistent object identity, scale, and motion across modalities and viewpoints (Fig. 4 (a)). Beyond the standard camera rig, AnyScene enables novel-view synthesis under flexible camera configurations at inference time without retraining (Fig. 4 (b)). AnyScene also supports fine-grained attribute control through user prompts. Specifying conditions such as weather or lighting in the text prompt (e.g., "cloudy", "nighttime") produces videos that faithfully reflect the desired attributes while preserving the geometry defined by occupancy. These capabilities highlight AnyScene as a flexible, high-quality generator suitable for diverse simulation and data synthesis tasks.

**Occupancy VAE and Occupancy Generation.** We first evaluate the BEV-based occupancy VAE, whose reconstruction quality determines the representation available to the occupancy generator. As shown in Tab. 1, reconstruction improves with latent resolution, and our VAE achieves the best results at $32 \times 128 \times 128$. At this matched latent shape, it outperforms $\mathcal{X}$-Scene (Yang et al., 2025c) by 8.80 mIoU points and 19.31 IoU points. We then evaluate conditional occupancy generation in Appendix Tab. 6. Against ground-truth occupancy, AnyScene achieves $19.01/15.58$ mIoU/IoU in 3D and $35.04/60.53$ after BEV projection, showing that horizontal scene structure is recovered more accurately than fine-grained vertical geometry. The BEV-to-layout scores of $52.45/68.97$ directly measure conditioning fidelity and confirm that the generated occupancy follows the input layout. Camera-view renderings obtain lower FID/KID than BEV renderings ($15.03/13.00$ versus $54.52/41.30$), consistent with perspective views being less sensitive to vertical errors. Appendix Fig. 11 provides qualitative results. The generated occupancy follows the specified road and agent layouts while completing coherent scene structures even where the LiDAR-derived ground truth is sparse.

**Video Generation Results.** Existing baselines do not support the full view-expansion setting, so we compare video generation under the common reference-view-conditioned protocol used by prior driving video methods. Following GenieDrive (Yang et al., 2025d), we re-evaluate all methods, in-




Figure 5: **Qualitative results of driving video generation.** Our method excels in object structure quality and appearance details.

Table 3: **Occupancy-generator ablation.** Ours uses teacher forcing, depth $L = 24$, and token concatenation.

| Variant | mIoU ↑ | IoU ↑ |
|---|---|---|
| **Ours** | **19.01** | **15.58** |
| w/o. Teacher-forcing | 3.23 | 0.98 |
| Depth $L = 16$ | 12.22 | 10.11 |
| Additive BEV | 13.59 | 7.60 |

Table 4: **View-expansion conditioning ablation**, built up from an unconditioned base.

| Configuration | FVD ↓ | mIoU ↑ | mAP ↑ |
|---|---|---|---|
| Base, no buffer | 112.20 | 32.96 | 17.70 |
| + Semantic buffer | 52.59 | 37.40 | 27.53 |
| + Coordinate buffer | 47.05 | 36.76 | 26.41 |
| + Both (**Ours**) | **43.19** | **38.26** | **27.73** |
| GT video | — | 41.03 | 36.38 |

cluding the action-conditioned OmniNWM (Li et al., 2026), using the same toolbox and protocol: FVD (Unterthiner et al., 2019) measures video quality, while mIoU and mAP measure controllability by applying BEVFormer (Li et al., 2024) to generated videos and comparing its predictions with ground-truth annotations. As shown in Tab. 2, our method achieves the best FVD at every evaluated length, with 35.98, 43.19, and 57.57 for 8, 16, and 49 frames, respectively. It also reaches 38.26 mIoU and 27.73 mAP, exceeding GenieDrive (Yang et al., 2025d) by 7.29 and 8.56 points. These results demonstrate strong temporal quality together with accurate semantic and spatial control. Figure 5 complements the quantitative comparison with generated object structures and appearance details. Since existing baselines do not support reference-free view expansion with flexible camera configurations, we evaluate this capability qualitatively in Fig. 4 and Appendix Figs. 18–19, which show temporally and cross-view consistent generation under flexible camera configurations and dynamic camera changes at inference time.

**Counterfactual Driving Video Generation.** Beyond reproducing recorded scenes, AnyScene generates counterfactual scenarios from user-specified BEV layout sequences. By modifying road geometry and agent states over time, these sequences can prescribe alternative traffic flows and scene configurations without corresponding real-world recordings. The occupancy generator realizes each sequence as temporally coherent 3D geometry, while the view-expansion module renders matching multi-view videos under flexible camera configurations. The same occupancy-centric representation also supports local occupancy edits and ego-trajectory perturbations without retraining. As shown in Fig. 6, shifting the ego trajectory changes the requested content while preserving the surrounding scene structure. Appendix Fig. 15 further pairs an arbitrary street-view image with a user-designed BEV layout to synthesize counterfactual observations, while Figs. 16 and 17 show counterfactual ego trajectories on unseen WOMD layouts and generation after removing vehicle occupancy.

**Densifying Sparse-View 3D Reconstruction.** Beyond video quality, the expanded rigs yield a direct and measurable benefit for sparse-view 3D reconstruction, where the 6 nuScenes cameras leave

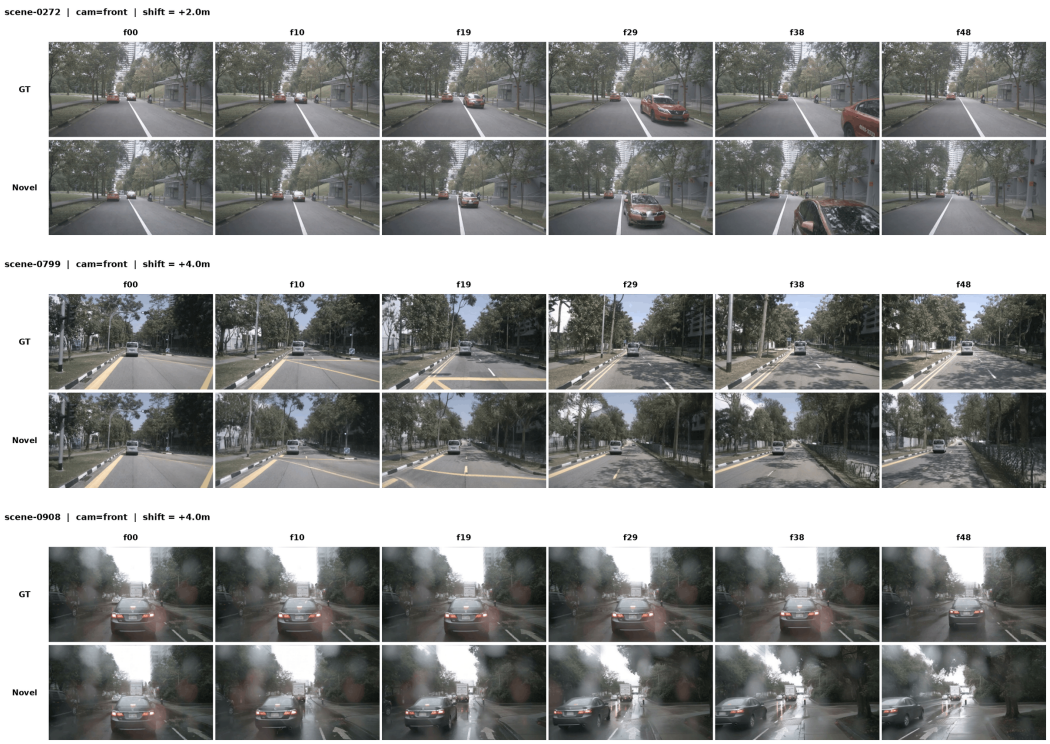


Figure 6: **Counterfactual scenario generation.** Each block re-renders a recorded nuScenes scene after laterally shifting the ego path by the offset in its title ($+2.0$, $+4.0$, and $+4.0$ m). Top row: the recorded video. Bottom row: our generation at the same timesteps. The requested viewpoint change is realized while the surrounding road structure and agents remain coherent. Front view shown.

Table 5: **Sparse-view 3D reconstruction with generated views.** 48 nuScenes validation scenes $\times$ 13 snapshots, metrics against LiDAR within 50 m (Acc / Comp in metres); Acc is, for each predicted point, the distance to the nearest LiDAR point, averaged. Starting from the 6 real cameras, we add 6 or 18 novel cameras with poses interpolated between adjacent real ones and images generated by the view-expansion module. Novel views improve completeness, recall, and coverage for both feed-forward models, while Acc stays within 0.03 m of the real-camera baseline.

| Views | Model | Comp↓ | Rec@0.5↑ | BEV Cov.↑ | F@0.5↑ | Acc↓ |
|---|---|---|---|---|---|---|
| 6 real | | 1.097 | 0.561 | 0.564 | 0.559 | 0.882 |
| 6 real +6 novel | Depth Anything 3 | 0.792 | 0.634 | 0.643 | 0.589 | 0.855 |
| 6 real +18 novel | | **0.678** | **0.679** | **0.675** | **0.604** | **0.851** |
| 6 real | | 0.940 | 0.598 | 0.568 | 0.613 | 0.707 |
| 6 real +6 novel | SurroundNEXO | 0.745 | 0.659 | 0.631 | 0.636 | 0.728 |
| 6 real +18 novel | | **0.643** | **0.700** | **0.664** | **0.657** | **0.701** |

wide unobserved wedges between adjacent views. We select 48 nuScenes validation scenes stratified by location and by day / rain / night, with 13 snapshots per scene (624 in total). Starting from the 6 real cameras, we add 6 or 18 novel cameras whose poses are interpolated between adjacent real ones and whose images are outpainted by the view-expansion module from the same occupancy. Following the calibrated-rig feed-forward reconstruction setup described in the blog (Jang, 2026), we jointly reconstruct all views of each snapshot using two pose-conditioned models with different inductive biases, Depth Anything 3 (Lin et al., 2025) and SurroundNEXO (Yuan et al., 2026). The predicted depth is back-projected to a point cloud, scale-aligned to LiDAR on the real views, and compared against LiDAR within 50 m; the full protocol is given in the supplementary material. As reported in Tab. 5, the novel views improve every completeness-type metric by a wide margin while

leaving accuracy on par with the real-camera baseline. With Depth Anything 3, the completeness error drops from $1.10$ to $0.68$ m, Rec@0.5 rises from $0.56$ to $0.68$, and BEV coverage from $0.56$ to $0.68$. SurroundNEXO reproduces the same picture, so the gain is a property of the generated views rather than of a particular reconstruction model. Across both reconstruction models, accuracy changes by at most $0.03$ m, while completeness, recall, and BEV coverage consistently improve as novel views are added. These results show that the generated views expand the reconstructed scene coverage while preserving the geometric accuracy of the original sparse-view reconstruction.

**Additional Qualitative Results.** The supplementary material provides additional visual evidence for the capabilities evaluated above. Appendix Fig. 11 shows layout-aligned occupancy generation; Figs. 12–14 examine temporal control and cross-view consistency; Figs. 15, 16, and 17 present counterfactual observation generation from an arbitrary street-view image and custom BEV layout, ego trajectories on unseen layouts, and occupancy-level editing; and Figs. 18 and 19 demonstrate surround-view synthesis with an optional reference image and expansion to flexible camera rigs.

### 3.4 Ablation Studies

**Effect of Occupancy Generator Design.** Table 3 evaluates three design choices against the full occupancy generator, which achieves $19.01$ mIoU and $15.58$ IoU. Removing teacher forcing causes the largest degradation, reducing the scores to $3.23/0.98$, which highlights the importance of clean historical context for learning stable autoregressive dynamics. Reducing the backbone depth from $L = 24$ to $L = 16$ lowers mIoU and IoU by $6.79$ and $5.47$ points, respectively, showing that model capacity benefits both semantic classification and occupied-space recovery. Replacing token concatenation with additive BEV fusion reduces mIoU and IoU by $5.42$ and $7.98$ points, respectively, showing that explicit interaction between BEV and occupancy tokens is important for semantic and geometric alignment.

**Effect of View-Expansion Module Design.** We build the conditioning up from an unconditioned base in Tab. 4, reporting FVD on 16-frame clips together with the mAP and mIoU obtained by running BEVFormer on the generated videos. The base reaches 112.20 FVD and 17.70 mAP. Either buffer alone recovers most of the headroom, the semantic buffer gaining 9.83 mAP and 59.61 FVD and the coordinate buffer 8.71 mAP and 65.15 FVD, a split that follows what each signal encodes: semantics favor per-frame recognizability, coordinates favor temporal fidelity. Adding the second buffer on top of the better single one is worth a further 0.20 mAP, 0.86 mIoU and 3.86 FVD, so the two signals overlap heavily and leave only a small complementary residual. Running the same perception model on real videos yields 36.38 mAP and 41.03 mIoU, bounding what generated data can reach under this protocol.

## 4 Conclusion and Future Work

In this paper, we presented AnyScene, a unified occupancy-centric framework for scalable and controllable driving data generation. Given diverse BEV layout sequences, AnyScene autoregressively generates temporally coherent semantic occupancy and uses the resulting geometry to synthesize consistent multi-view videos without real reference frames and with flexible camera counts and poses. To support high-frequency generation, we built *nuCraftv2*, a 12 Hz dataset with synchronized BEV layouts and high-quality dense semantic occupancy. Experiments show that the occupancy generator generalizes zero-shot across user-designed layouts, unseen datasets, and global maps, while the video generator achieves state-of-the-art performance. We further demonstrate its application to counterfactual scenario generation, including alternative traffic flows, occupancy edits, and ego-trajectory perturbations, as well as densifying sparse-view 3D reconstruction while preserving geometric accuracy.

**Limitations and Future Work.** The current framework realizes traffic flows specified by BEV layout sequences but does not generate traffic behavior or model reactive interactions; agents follow externally provided layouts rather than responding to one another or to the ego vehicle. Integrating AnyScene with a traffic behavior model could enable reactive agent synthesis, while deployment in a closed-loop simulation platform could support targeted generation of corner cases for autonomous-driving training and evaluation.

## AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools solely for language polishing, including improving writing clarity and correcting grammar. We have reviewed all AI-assisted work and verified the resulting edits. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. In *ICLR*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023b.

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. *CVPR*, 2020.

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

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.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025a.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo De Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al. Omnire: Omni urban scene reconstruction. In *International Conference on Learning Representations*, 2025b.

MMEngine Contributors. MMEngine: Openmmlab foundational library for training deep learning models. 2022.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9710–9719, 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. In *International Conference on Learning Representations*, volume 2024, pp. 22841–22860, 2024a.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28135–28144, 2025.

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. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024b.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023a.

Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructing world model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024.

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, 2023b.

Binyuan Huang, Yuqing Wen, Yucheng Zhao, Yaosi Hu, Yingfei Liu, Fan Jia, Weixin Mao, Tiancai Wang, Chi Zhang, Chang Wen Chen, et al. Subjectdrive: Scaling generative data in autonomous driving via subject control. In *AAAI*, volume 39, pp. 3617–3625, 2025.

Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4369–4379, 2023a.

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

Wonbong Jang. Feed-forward 3D reconstruction meets a real AV camera rig. `https://wbjang.github.io/blog/posts/av-rig-recon/`, August 2026. Blog post, accessed September 22, 2026.

Junpeng Jiang, Gangyi Hong, Miao Zhang, Hengtong Hu, Kun Zhan, Rui Shao, and Liqiang Nie. Dive: Efficient multi-view driving scenes generation based on video diffusion transformer. *arXiv preprint arXiv:2504.19614*, 2025a.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025b.

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

Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, and Sung-Eui Yoon. Semcity: Semantic scene generation with triplane diffusion. In *CVPR*, pp. 28337–28347, 2024.

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. In *CVPR*, 2025.

Bohan Li, Zhuang Ma, Dalong Du, Baorui Peng, Zhujin Liang, Zhenqiang Liu, Xianda Guo, Zheng Zhu, Chao Ma, Yueming Jin, Xin Jin, Hao Zhao, and Wenjun Zeng. Omninwm: Omniscient driving navigation world models. In *European Conference on Computer Vision*, 2026.

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*, 47(3):2020–2036, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv:1708.02002*, 2017.

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*, 2024a.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *ECCV*, 2024b.

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. In *ICCV*, 2025.

OpenStreetMap contributors. OpenStreetMap. `https://www.openstreetmap.org`, 2025. Accessed: May 15, 2025.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pp. 2446–2454, 2020.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36:64318–64330, 2023.

Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *CVPR*, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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

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.

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

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.

Francis Williams, Jiahui Huang, Jonathan Swartz, Gergely Klar, Vijay Thakkar, Matthew Cong, Xuanchi Ren, Ruilong Li, Clement Fuji-Tsang, Sanja Fidler, et al. fvdb: A deep-learning framework for sparse, large scale, and high performance spatial intelligence. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets and Benchmarks Track*, 2021.

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

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.

Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18268–18278, 2023.

Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pp. 156–173. Springer, 2024.

Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023a.

Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiangyu Yue, and Li Chen. Resim: Reliable world simulation for autonomous driving. *Advances in Neural Information Processing Systems*, 38:167710–167741, 2025a.

Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26933–26943, 2025b.

Yu Yang, Alan Liang, Jianbiao Mei, Yukai Ma, Yong Liu, and Gim Hee Lee. X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. *Advances in Neural Information Processing Systems*, 38:104415–104451, 2025c.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1389–1399, 2023b.

Zhenya Yang, Zhe Liu, Yuxiang Lu, Liping Hou, Chenxuan Miao, Siyi Peng, Bailan Feng, Xiang Bai, and Hengshuang Zhao. Geniedrive: Towards physics-aware driving world model with 4d occupancy guided video generation. *arXiv preprint arXiv:2512.12751*, 2025d.

Shuai Yuan, Runxi Tang, Yuzhou Ji, Fudong Ge, Hanshi Wang, Yifei Wang, Xianming Zeng, Jianyun Xu, Xingliang Liu, Yanfeng Wang, and Zhipeng Zhang. Surroundnexo: Ego-centric metric bridging for spatially consistent geometry in autonomous driving. *arXiv preprint arXiv:2606.16960*, 2026.

Junge Zhang, Qihang Zhang, Li Zhang, Ramana Rao Kompella, Gaowen Liu, and Bolei Zhou. Urban scene diffusion through semantic occupancy map. *arXiv preprint arXiv:2403.11697*, 2024.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *ICCV*, 2025.

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.

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

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *ECCV*, 2024.

Xingguang Zhong, Yue Pan, Jens Behley, and Cyrill Stachniss. Shine-mapping: Large-scale 3d mapping using sparse hierarchical implicit neural representations. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8371–8377. IEEE, 2023.

Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *CVPR*, pp. 21634–21643, 2024.

Benjin Zhu, Zhe Wang, and Hongsheng Li. nucraft: Crafting high resolution 3d semantic occupancy for unified 3d scene understanding. In *European Conference on Computer Vision*, pp. 125–141. Springer, 2024.

Benjin Zhu, Xiaogang Wang, and Hongsheng Li. Consistentcity: Semantic flow-guided occupancy dit for temporally consistent driving scene synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26382–26392, 2025.

# A Related Works

## A.1 Controllable Occupancy Generation

Semantic occupancy has emerged as a structured volumetric representation that encodes both geometry and semantics, making it suitable for controllable scene generation. Early efforts on layout-conditioned occupancy synthesis include UrbanDiffusion (Zhang et al., 2024), which generates semantic occupancy from BEV maps using a 3D diffusion model. More recent works have further explored controllable generation from BEV or layout conditions, such as UniScene (Li et al., 2025), which proposes an occupancy-centric pipeline for generating semantic occupancy from BEV layouts; InfiniCube (Lu et al., 2025), which enables unbounded and controllable dynamic 3D scene generation; WoVoGen (Lu et al., 2024b), which develops a world volume-aware diffusion model for controllable multi-camera driving scene generation; X-Scene (Yang et al., 2025c), which advances large-scale generation with flexible multi-granular control; and ConsistentCity (Zhu et al., 2025), which improves temporal consistency under BEV conditioning via a semantic flow-guided Occupancy DiT.

However, most existing methods still struggle to achieve strong controllability from diverse or user-defined BEV inputs. Our work addresses this limitation by developing an autoregressive occupancy generator that enables precise and high-frequency control from flexible BEV layouts.

## A.2 Controllable Video Generation

The advent of general-purpose video diffusion models (Blattmann et al., 2023b;a) has spurred significant progress in driving-specific video generation. Early single-view world models (Kim et al., 2021; Wang et al., 2023; Zheng et al., 2024) were constrained to short sequences, whereas more recent methods scale to longer, higher-resolution videos with enhanced trajectory and action conditioning (Gao et al., 2024b; Hu et al., 2024; Zhang et al., 2025; Hu et al., 2023a; Russell et al., 2025). A parallel line of research targets *multi-view* driving video generation conditioned on BEV layouts, HD maps, or 3D bounding boxes, exemplified by MagicDrive (Gao et al., 2024a) and its long-video extension MagicDrive-V2 (Gao et al., 2025), as well as Panacea (Wen et al., 2024), DriveDreamer-2 (Zhao et al., 2024), SubjectDrive (Huang et al., 2025), UniMLVG (Chen et al., 2024), CogDriving (Lu et al., 2024a), DiVE (Jiang et al., 2025a), and Glad (Xie et al., 2025). To improve cross-view coherence, existing methods commonly employ lightweight cross-view attention mechanisms for information exchange among camera views. Occupancy-guided pipelines (Lu et al., 2024b; Li et al., 2025; Lu et al., 2025) further introduce 3D structural priors by rendering occupancy representations into semantic or depth maps as conditioning signals for video generation. However, these methods still rely on indirect structural guidance or limited cross-view interactions, which makes it challenging to enforce precise geometric correspondence across cameras. In addition, jointly generating multiple views remains computationally expensive, limiting their efficiency and scalability.

## A.3 Generative Simulation for Autonomous Driving

Scalable and realistic simulation is fundamental to the development and validation of autonomous driving systems. Traditional reconstruction-based simulators, which rely on neural radiance fields (Yang et al., 2023b) or 3D Gaussian Splatting (Yan et al., 2024; Chen et al., 2025b), can achieve high visual fidelity when replaying logged trajectories. However, their per-scene optimization inherently limits scenario diversity.

To overcome this limitation, recent research has increasingly focused on generative approaches capable of synthesizing novel driving scenarios. Early generative simulators primarily targeted video synthesis, such as DriveDreamer-2 (Zhao et al., 2024) and the GAIA series (Hu et al., 2023a; Russell et al., 2025), which employ video diffusion models to generate long-horizon driving videos under trajectory or action conditioning. To improve geometric consistency and multi-modal alignment, more recent works have explored unified frameworks that jointly generate multiple data modalities. UniScene (Li et al., 2025) introduces an occupancy-centric pipeline that first generates semantic occupancy from BEV layouts and subsequently produces aligned multi-view videos and LiDAR. X-Scene (Yang et al., 2025c) and Cosmos-Drive-Dreams (Ren et al., 2025) further advance large-scale multi-modal generation by improving spatio-temporal consistency. DriveArena (Yang et al.,

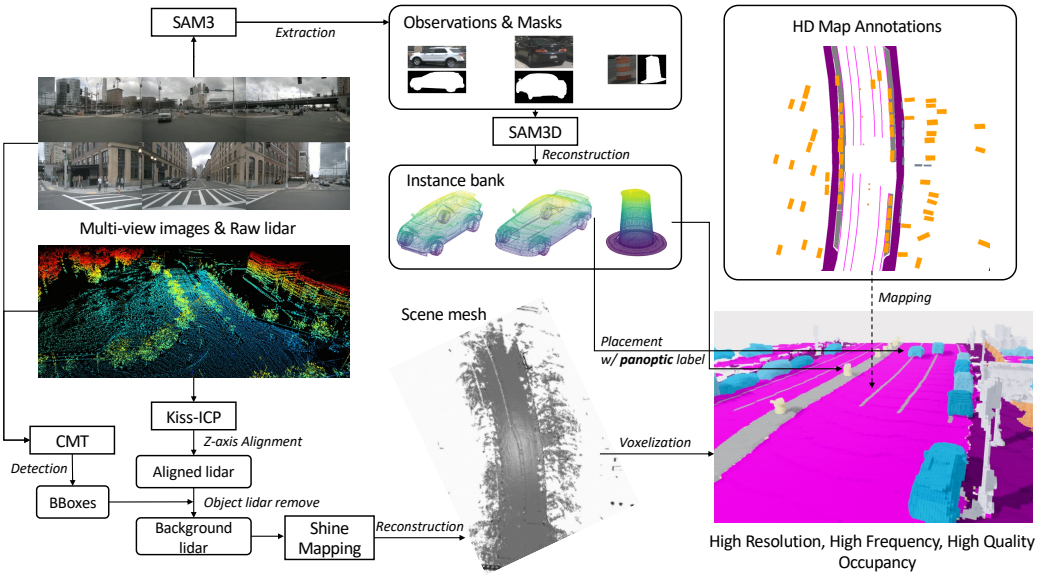


Figure 7: The curation pipeline of the nuCraftv2 dataset.

2025b) and ReSim (Yang et al., 2025a) explore controllable and large-scale generative simulation using autoregressive models and heterogeneous training data.

These works reflect a growing trend toward controllable and multi-modal generative simulation. In contrast to these prior efforts, our framework positions forecasted occupancy as the central representation and integrates autoregressive occupancy generation with geometry-grounded view expansion to enable high-fidelity and flexible scene synthesis from diverse BEV layouts.

## B NUCRAFTV2 DATA CURATION

nuCraftv2 builds upon the high-resolution occupancy formulation of nuCraft (Zhu et al., 2024). It reconstructs dense static geometry from aggregated LiDAR sweeps, models dynamic agents with per-instance assets, and assembles both components at synchronized 12 Hz timestamps using object poses and HD-map priors. The resulting dataset provides dense, instance-aware occupancy at voxel sizes of $0.1/0.2/0.4\,\mathrm{m}$. Figure 7 summarizes the curation pipeline.

**Dynamic-instance priors.** nuScenes provides ground-truth 3D boxes only at the 2 Hz key-frame rate. We therefore run a cross-modal LiDAR–camera detector (Yan et al., 2023) over every sweep to obtain dense detections for constructing the 12 Hz sequence and removing moving points from the static reconstruction. For each annotated foreground instance, we select its best observation across the six surround cameras according to visibility and projected area. SAM3 (Carion et al., 2025) extracts the object mask, and SAM3D (Chen et al., 2025a) reconstructs a canonical 3D asset that is reused across frames. Canonical in-box LiDAR points provide a fallback for instances whose reconstruction is insufficient. Examples are shown in Fig. 8.

**Static-scene reconstruction.** We refine the per-frame poses with KISS-ICP to recover vertical motion and fine $\mathrm{SE}(3)$ alignment. Points inside ground-truth or detected foreground boxes are removed to form a clean static background. Key-frame panoptic labels come from nuScenes, while nearby non-key frames receive labels through $k{=}3$ KNN propagation in the global frame. The cleaned sweeps and refined poses are then processed by SHINE-Mapping (Zhong et al., 2023), which reconstructs a semantic static-scene mesh for subsequent assembly.

**Per-frame occupancy assembly.** At each synchronized timestamp, we crop the static reconstruction to the $[-51.2, 51.2]^2 \times [-5, 5]\,\mathrm{m}$ region around the ego vehicle. Each foreground asset is scaled and aligned to its annotated box, with sparse instances densified by NKSR (Huang et al., 2023a). We

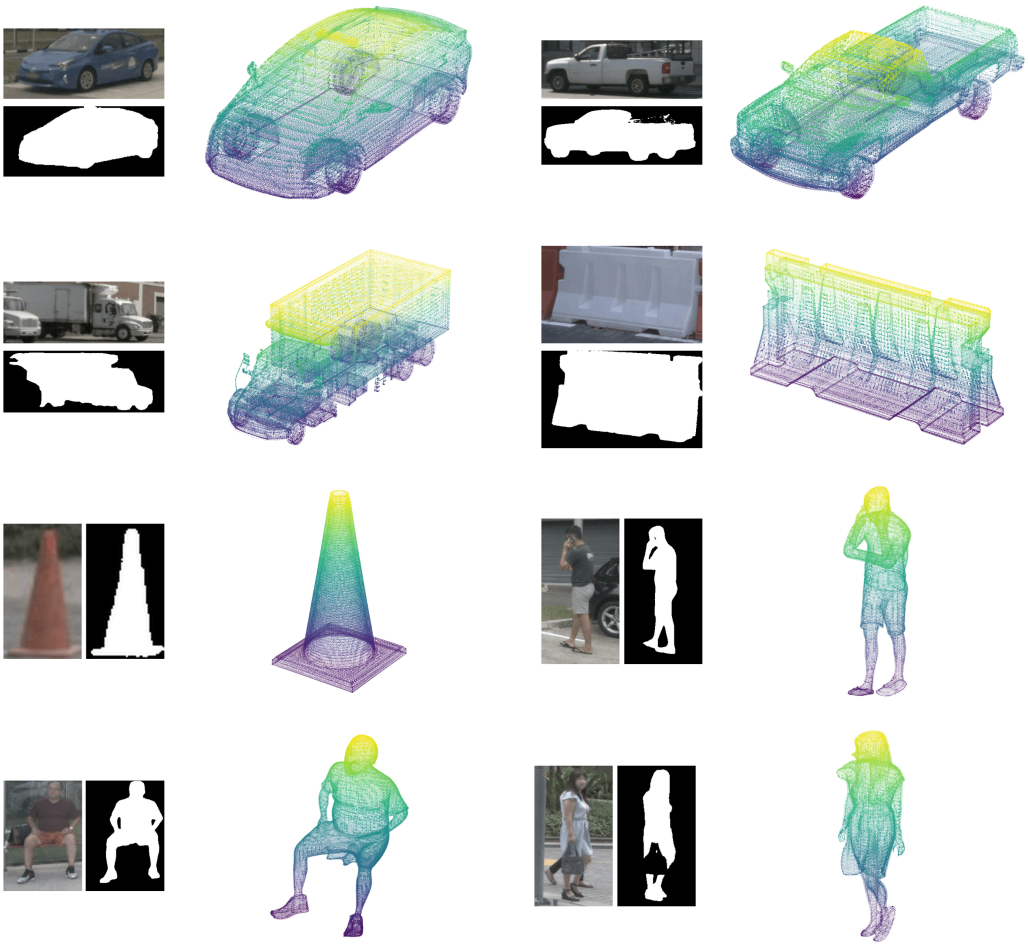

Figure 8: Instance extraction and reconstruction using SAM3 (Carion et al., 2025) and SAM3D (Chen et al., 2025a).

voxelize the union of static and dynamic points and refine road structures using the HD-map layout, producing the final panoptic occupancy. Figures 9 and 10 compare the resulting occupancy with existing representations.

### B.1 Experimental Representation

BEV layouts are constructed from the nuScenes HD map and 3D bounding-box annotations. Each layout is rasterized at $0.4\,\mathrm{m}$ on a $256\times256$ grid with 15 multi-hot channels representing 10 dynamic-agent classes and 5 static-map primitives. All occupancy-generation experiments use a volume spanning $[-51.2, 51.2]^2 \times [-5.0, 5.0]\,\mathrm{m}$ around the ego vehicle with a voxel size of $0.4\,\mathrm{m}$. This produces a $256 \times 256 \times 25$ tensor over 21 semantic classes, including free space. We use the original nuScenes multi-view camera images at $12\,\mathrm{Hz}$.

## C Additional Implementation Details

### C.1 BEV Layout Interface

A BEV layout is a multi-hot raster with interpretable channels. HD-map annotations and 3D boxes from different driving datasets can be projected into this shared representation, while user edits directly specify changes to roads or agents. By separating scene structure in the BEV layout from

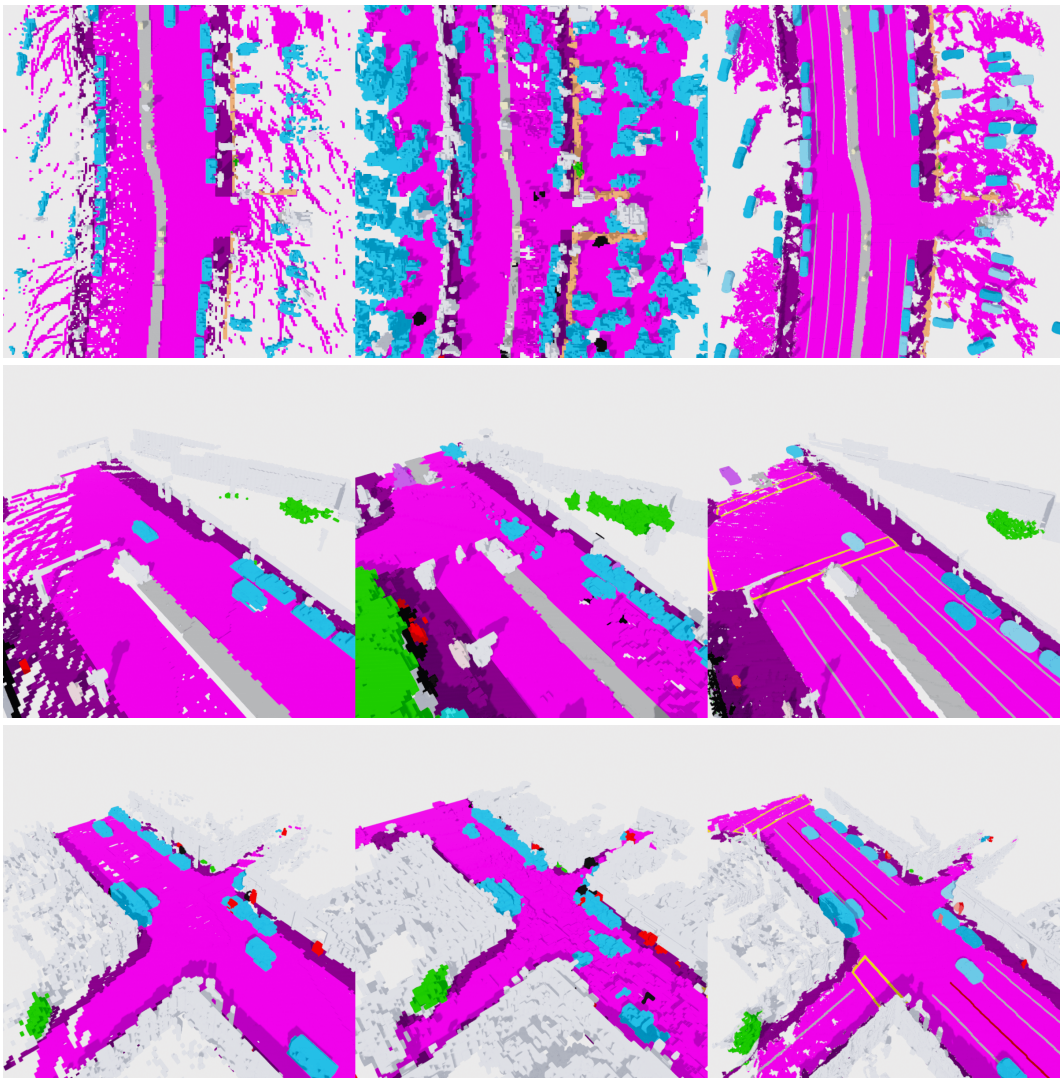

Figure 9: Qualitative comparison of occupancy representations. From left to right: OccWorld (2 Hz) (Zheng et al., 2023), UniScene (12 Hz) (Li et al., 2025), and nuCraftv2 (12 Hz).

occupancy and video appearance, the same interface supports unseen-dataset and user-designed layouts.

### C.2 BEV-Based Occupancy VAE

Given a 3D occupancy grid $\mathcal{O}_t \in \mathbb{R}^{H\times W\times Z}$, we convert each voxel label into a learnable $C'$-dimensional embedding and concatenate the embeddings along the height axis at each BEV location (Zheng et al., 2023). A convolutional and axial-attention encoder maps the resulting 2D tensor to a latent $\mathbf{z}_t \in \mathbb{R}^{h\times w\times C_z}$, and a symmetric 2D decoder reconstructs $\widehat{\mathcal{O}}_t$, as shown in Fig. 3(a). This representation avoids volumetric processing and remains compatible with image and video diffusion backbones (Zhang et al., 2024; Bian et al., 2025; Yang et al., 2025d; Wan et al., 2025).

We train the VAE using focal loss (Lin et al., 2017), Lovasz-Softmax loss, and KL regularization:

$$\mathcal{L}_{\text{occ}}^{\text{vae}} = \mathcal{L}_{\text{focal}} + \lambda_1 \mathcal{L}_{\text{LS}} + \lambda_2 \mathcal{L}_{\text{KL}}. \tag{2}$$

After training, the encoder and decoder are frozen and occupancy sequences are modeled in the latent space.

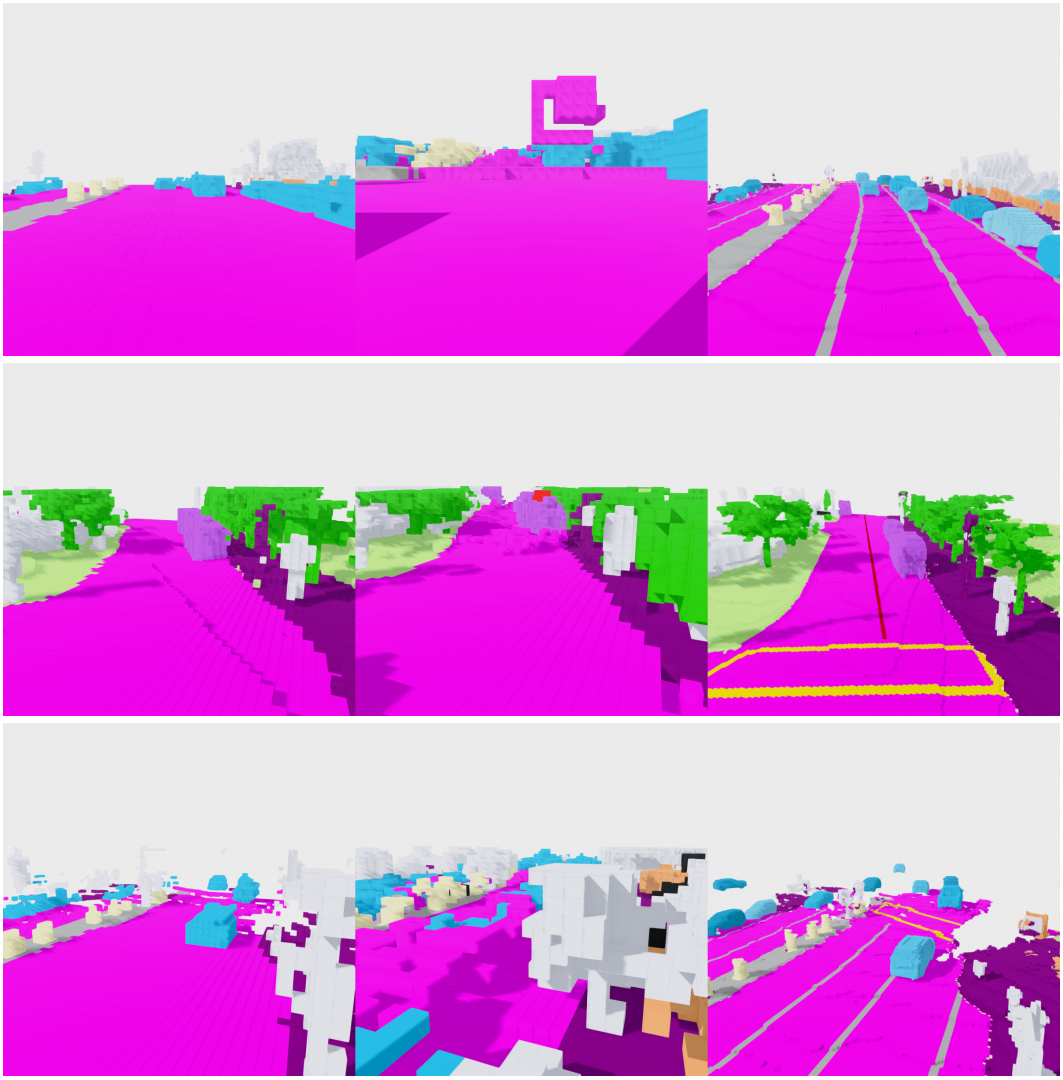

Figure 10: Additional qualitative comparison of occupancy representations. From left to right: OccWorld (2 Hz) (Zheng et al., 2023), UniScene (12 Hz) (Li et al., 2025), and nuCraftv2 (12 Hz).

### C.3 BEV-Conditioned Occupancy Generator

**Architecture and tokenization.** Each active class in the multi-hot BEV layout is represented by a learnable embedding; active embeddings are summed at each location, while empty locations use a separate embedding. The occupancy generator contains $L$=24 spatio-temporal transformer blocks with hidden dimension $D$=768, 12 attention heads, and an FFN expansion factor of 4. The BEV encoder and occupancy patch embedding map both streams to a common $32 \times 32$ grid with 1024 tokens per frame. Training uses $T$=4 frames, while inference uses 30 Euler steps and classifier-free guidance with scale $w$=2.0.

**Dual-stream causal modeling.** Each block contains spatial self-attention, causal temporal self-attention, and a SwiGLU feed-forward network under AdaLN-Zero modulation. Spatial attention jointly processes BEV and noisy occupancy tokens. Temporal attention interleaves the clean and noisy streams under the mask in Fig. 3(b), allowing the current noisy frame to attend to clean preceding frames. At inference, these clean latents are replaced by the model's previously generated predictions, yielding autoregressive occupancy generation.

**Training procedure.** The occupancy generator is trained in two phases with AdamW and bfloat16 mixed precision. The first phase freezes the zero-gated temporal branch and learns BEV-conditioned single-frame generation. The second phase activates the temporal layers and fine-tunes four-frame sequences, applying a lower learning rate to the parameters inherited from the first phase. Gaussian noise augmentation and history dropout are applied to the clean stream during temporal training.

**Flow matching objective.** For each frame $t$, we sample a timestep $\tau_t$ from a logit-normal distribution and construct

$$\mathbf{z}_t^{\tau_t} = (1 - \tau_t)\mathbf{z}_t + \tau_t \boldsymbol{\epsilon}_t, \tag{3}$$

where $\boldsymbol{\epsilon}_t$ is standard Gaussian noise. The model predicts the velocity field using a weighted mean-squared-error objective, with layout-derived weights emphasizing small classes such as pedestrians and bicycles.

### C.4 Geometry-Grounded View Expansion

**Unified mode-mixture training.** We train a single video diffusion model with five conditioning modes over triplets of adjacent cameras. The first three generate one, two, or three target views from geometry, optionally using the first target frame as a reference. The remaining modes outpaint one target from either one neighboring anchor view or two outer anchor views.

| Mode | Anchor views | Target views | Target-frame reference |
|---|---|---|---|
| Single-view generation | 0 | 1 | Optional |
| Two-view generation | 0 | 2 | Optional |
| Three-view generation | 0 | 3 | Optional |
| One-anchor outpainting | 1 | 1 | – |
| Two-anchor outpainting | 2 | 1 | – |

All active-view latents are concatenated and jointly noised at a shared diffusion timestep. The control branch receives clean latents at anchor positions, zeros at target positions, and a binary mask indicating observed regions. The reconstruction loss down-weights anchor views so that target views dominate the gradient, while a curriculum gradually balances direct generation and outpainting modes.

**Autoregressive view expansion.** Starting from occupancy alone, the model first synthesizes the front-left, front, and front-right views jointly. It then outpaints the back-left and back-right views from their neighboring front views, followed by the rear view from the two generated back-side anchors. When a front-camera image is available at the first frame, it serves as an optional appearance reference without changing the remaining procedure. The same outpainting operation inserts virtual cameras between existing views; we demonstrate expansion from 6 to 12, 18, and 24 cameras without retraining.

**Architecture and training.** The module uses a frozen Wan2.1-T2V-14B backbone (Wan et al., 2025) with a trainable VACE-style control branch (Jiang et al., 2025b). Target-view RGB latents and Plücker camera embeddings are patchified into the control stream, while semantic and coordinate buffers are embedded and added as geometry conditions following InfiniCube (Lu et al., 2025). Eight control blocks inject residual hints at uniformly spaced layers of the 40-block backbone. Only the control branch is updated; the DiT backbone, text encoder, and VAE remain frozen. We train for 5k iterations at $480 \times 832$ resolution using 49 sampled frames, AdamW with learning rate $2 \times 10^{-5}$ and weight decay 0.01, bfloat16 mixed precision, and gradient checkpointing.

## D Evaluation Details

### D.1 Occupancy Generation Metrics

Inspired by X-Scene (Yang et al., 2025c) and UniScene (Li et al., 2025), we evaluate occupancy generation using the metrics in Tab. 6.

**3D Volume** accumulates a voxel-level confusion matrix over the validation split; mIoU averages the non-free semantic classes, while IoU measures occupied versus free space. **BEV Top-down** applies

Table 6: **Occupancy generation quality on nuCraftv2-val.** For FID and KID, scores are averaged across the depth and semantic streams.

| Method | 3D Volume (↑) | | BEV Top-down (↑) | | BEV vs. Layout (↑) | | Cam Render (↓) | | BEV Render (↓) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mIoU | IoU | mIoU | IoU | mIoU | IoU | FID | KID | FID | KID |
| AnyScene (ours) | **19.01** | **15.58** | **35.04** | **60.53** | **52.45** | **68.97** | **15.03** | **13.00** | **54.52** | **41.30** |

the same metrics after projecting both volumes to a 2D label map. **BEV vs. Layout** compares the top-down occupancy projection with the conditioning layout. **Cam Render** computes FID and KID using camera-view depth and semantic renderings, while **BEV Render** applies the same metrics to BEV semantic and height projections. All metrics are derived from the predicted occupancy and its corresponding ground truth.

### D.2 Sparse-View Reconstruction Protocol

We evaluate 48 nuScenes validation scenes stratified by location and day, rain, or night conditions, with 13 snapshots per scene and 624 snapshots in total. Starting from the six real cameras, we add 6 or 18 novel cameras whose poses are interpolated between adjacent real cameras and whose images are generated by the view-expansion module from the same occupancy. All views of each snapshot are jointly reconstructed using Depth Anything 3 (Lin et al., 2025) or SurroundNEXO (Yuan et al., 2026).

Predicted depth is back-projected to a point cloud, scale-aligned to LiDAR on the real views, and restricted to the top 70% confidence within $50\,\mathrm{m}$. Accuracy and completeness are point-to-LiDAR and LiDAR-to-point distances; Rec@0.5 and F@0.5 use a $0.5\,\mathrm{m}$ threshold; BEV coverage measures the fraction of LiDAR-occupied BEV cells reached by the reconstruction. As reported in Tab. 5, the generated views improve completeness-related metrics while maintaining reconstruction accuracy across both models. In a leave-one-camera-out test on eight scenes, adding 18 novel views raises the held-out image coverage from 0.17 to 0.93.

## E Additional Qualitative Results

### E.1 Occupancy Generation

Figure 11 compares the conditioning layout, ground-truth occupancy, and generated occupancy. The generated occupancy follows the specified road and agent layouts while maintaining coherent 3D scene structure.

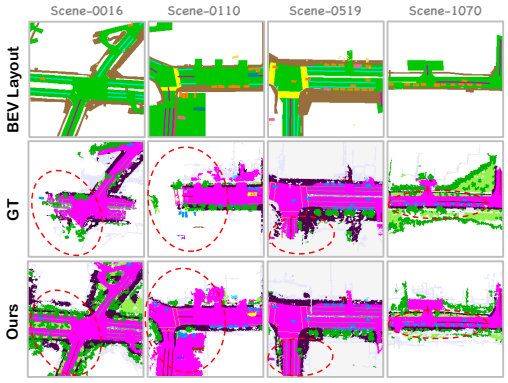


Figure 11: **Qualitative occupancy generation.** From top to bottom: BEV layout, ground-truth occupancy, and generated occupancy.

### E.2 Temporal and Cross-View Control

Figures 12–14 show that generated frames follow the per-frame semantic and coordinate buffers over both consecutive and longer sequences, while preserving consistent overlapping content across adjacent views.

### E.3 Counterfactual Editing

The occupancy-centric interface supports ego-trajectory perturbation and local scene editing. Beyond the lateral shifts on recorded nuScenes scenes in Fig. 6, Fig. 15 combines an arbitrary street-view image with a custom BEV layout to generate counterfactual observations, Fig. 16 generates alternative ego trajectories on BEV layouts from an unseen dataset, and Fig. 17 removes vehicle occupancy while preserving the surrounding scene.

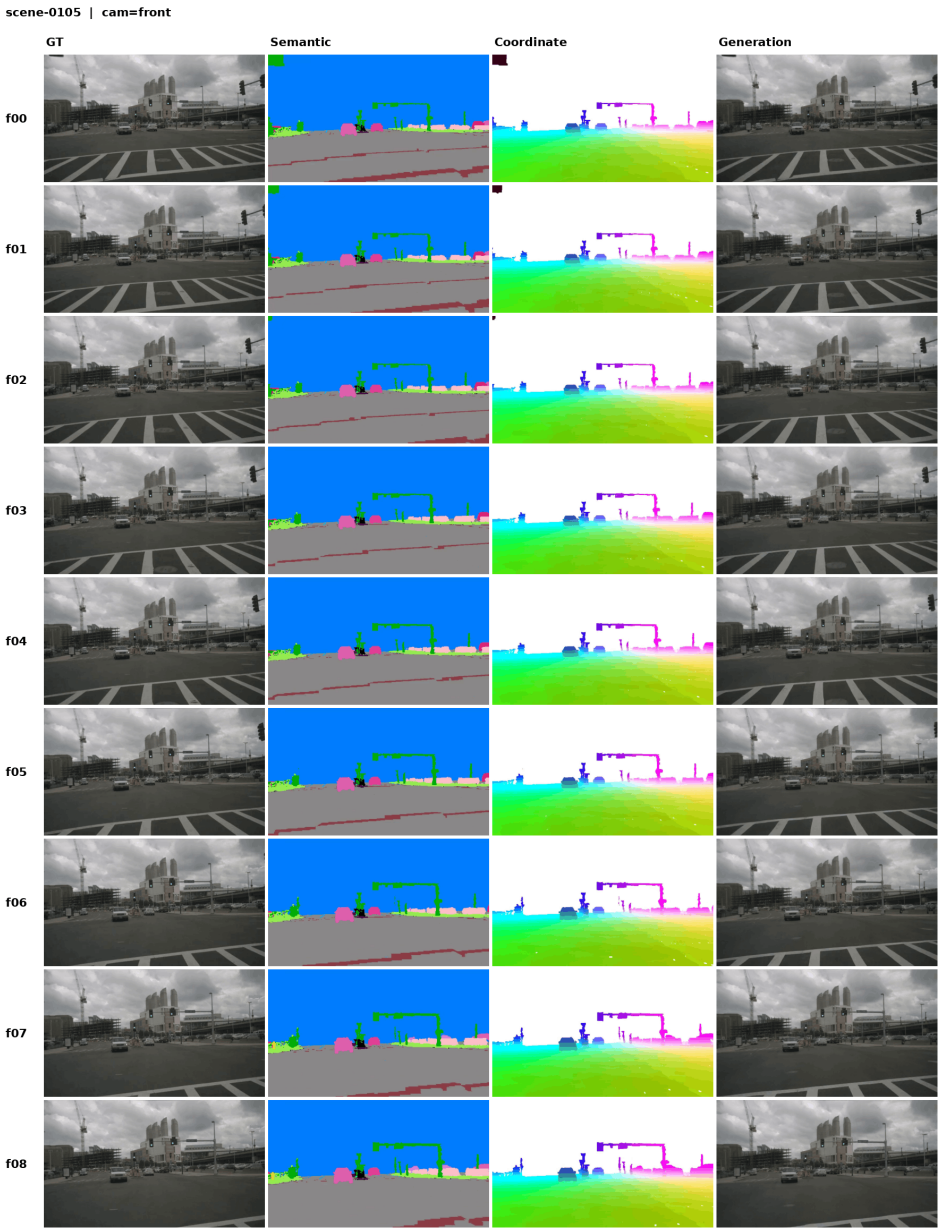


Figure 12: **Per-frame condition control over consecutive frames.** Each row shows the ground truth, semantic buffer, coordinate buffer, and generated frame.

### E.4 Flexible Camera Rigs

Figures 18 and 19 demonstrate surround-view generation from an optional front-camera reference and camera-rig expansion from the six baseline cameras to 12, 18, and 24 views.

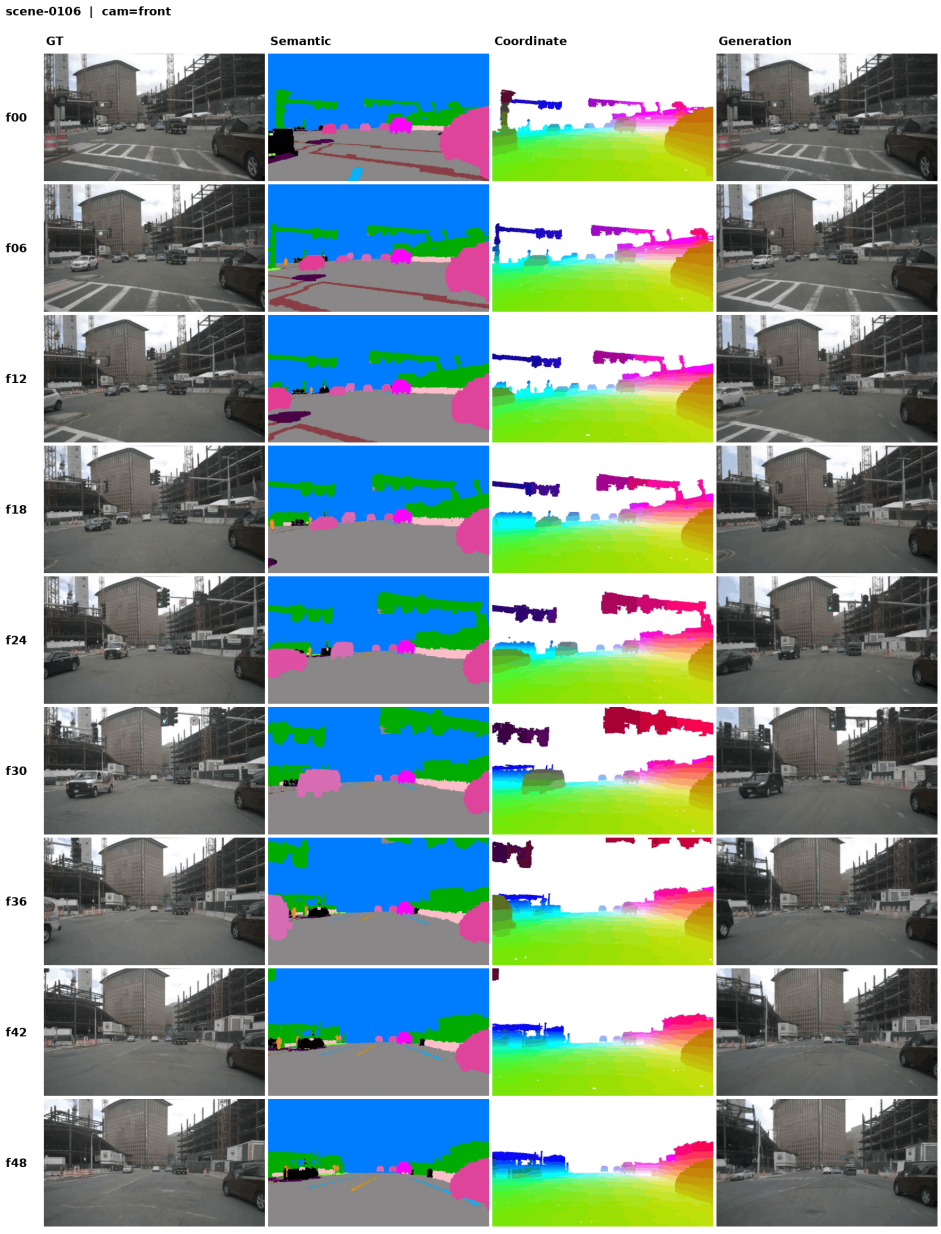


Figure 13: **Per-frame condition control.** Generated frames remain aligned with the 12 Hz geometric conditions.




Figure 14: **Cross-view consistency in two- and three-view outpainting.** Target views generated from geometric buffers and neighboring anchors preserve consistent overlapping content.

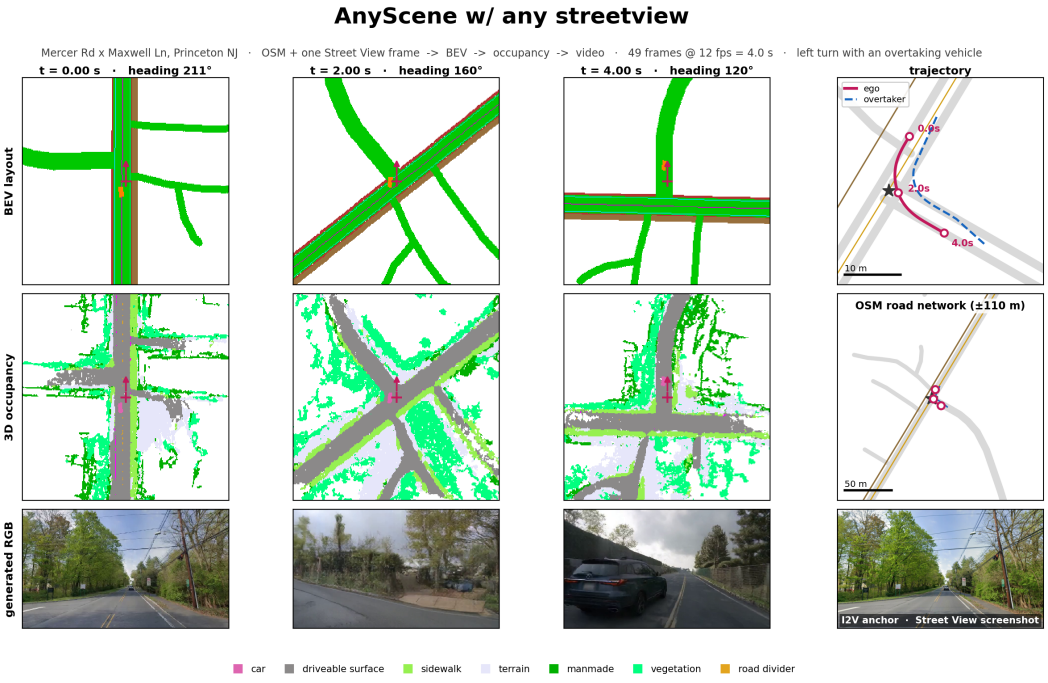


Figure 15: **Counterfactual observation generation from an arbitrary street-view image.** The street-view image provides appearance context, while a custom BEV layout specifies road geometry and ego and agent trajectories. AnyScene generates the corresponding occupancy sequence and RGB observations without matching driving footage.




Figure 16: **Counterfactual ego trajectories on unseen WOMD layouts.** AnyScene renders three laterally separated candidate ego trajectories (red / green / blue; minimum lateral separation 3.08–3.18 m) on Waymo Open Motion Dataset (Ettinger et al., 2021) scenes. Within each row the agent configuration and the text prompt are held fixed, so the ego path is the only variable. Left: the local BEV layout with the three paths and all agent boxes at the displayed timestep. Right: the corresponding generated frames. No real footage exists for these layouts, so no ground truth is shown. Front view shown.

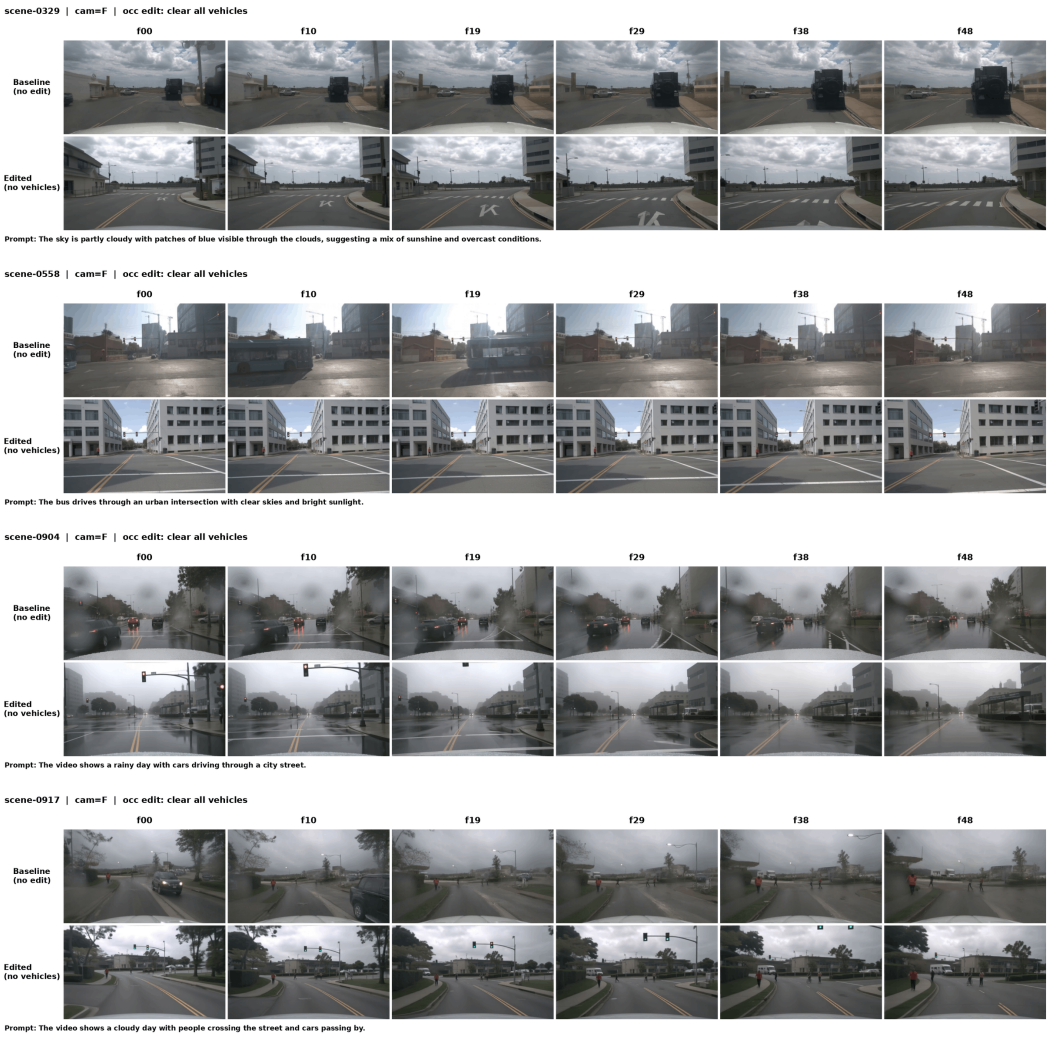


Figure 17: **Occupancy editing through vehicle clearing.** Removing vehicle voxels changes the requested content while preserving the surrounding static scene.




Figure 18: **Anchor-image surround generation.** Given the front-camera frame at the first timestep, the model synthesizes the complete six-view surround rig.

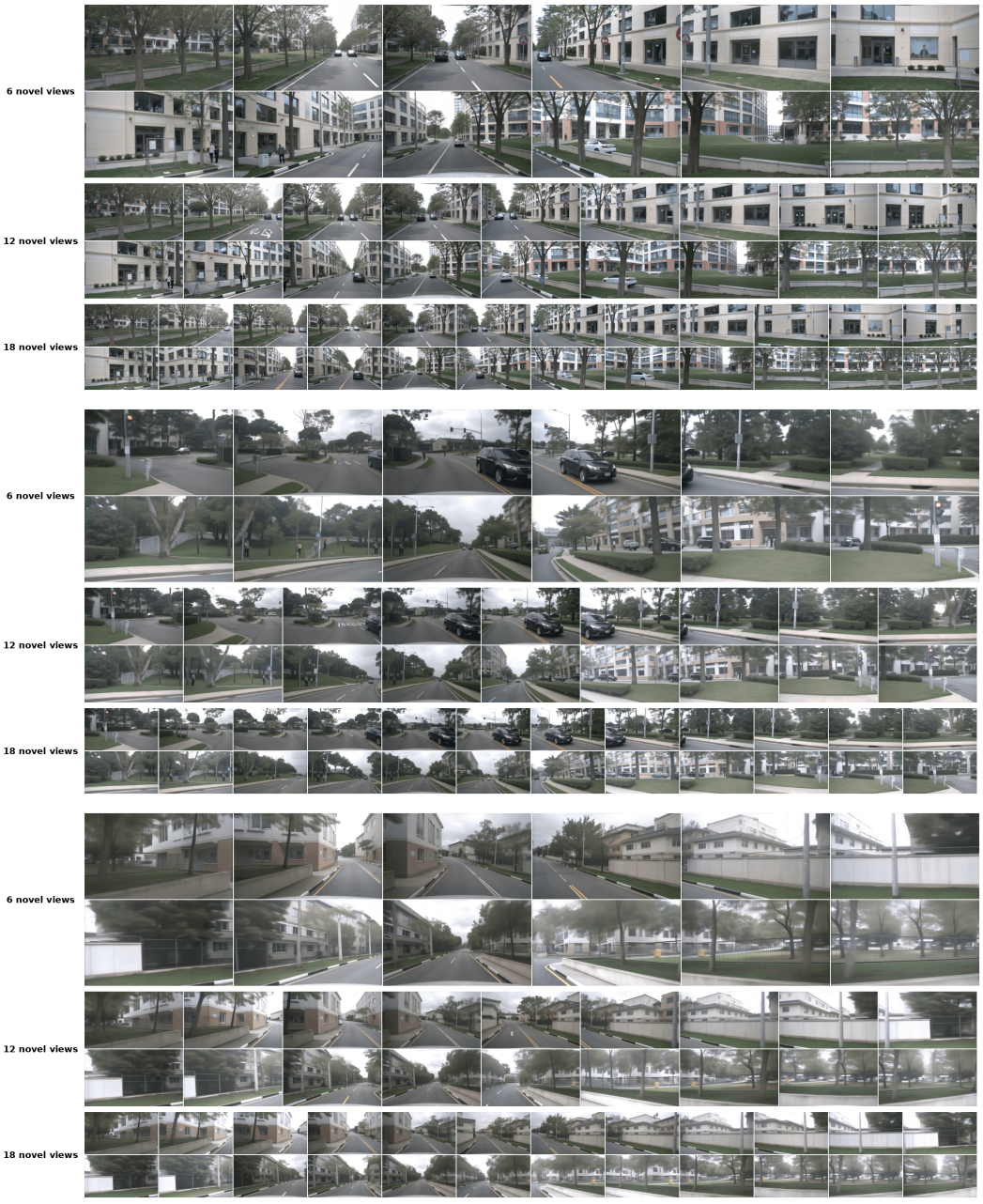


Figure 19: **Geometry-grounded view expansion.** Each example expands the six baseline cameras into 12-, 18-, and 24-view rigs.