SceneDirector: Bridging Explicit Geometry and Generative Priors for Unified Driving Scene Editing

Anonymous Authors[1]

Figure 1. SceneDirector is a unified framework for multi-view driving video editing. It enables simultaneous 3D box-defined **object editing** (insertion, deletion, replacement, repositioning) and **ego-trajectory editing** in a single inference pass. By bridging geometric guidance with generative priors, it reconciles photorealistic synthesis for object editing and structural consistency for trajectory control.

Abstract

Validating autonomous driving systems requires diverse scenarios, yet real-world data collection is biased and costly. Editing existing driving logs offers a scalable solution, but simultaneously editing objects and ego-trajectory—termed unified editing—remains challenging. Current methods face an inherent dilemma: generative flexibility for object editing and physical precision for trajectory control. To address this, we introduce SceneDirector, a diffusion-based framework that bridges explicit geometry and generative priors. For explicit geometry, we leverage LiDAR-guided depth completion to construct dense scene geometry and integrate editable 3D assets to form a Unified Geometric Scaffold, providing rigorous structural guidance for unified editing. To leverage generative priors, we encode the source video into a Static Texture Bank to provide rich appearance context. Our proposed Mask-Gated Reference Attention bridges these modalities. Guided by a geometric uncertainty metric, this mechanism dynamically regulates the interaction between the scaffold and the bank—preserving reliable geometry while adaptively injecting textures for semantic refinement. Extensive evaluations demonstrate that SceneDirector outperforms state-of-the-art methods in both controllability and visual quality.

1. Introduction

Autonomous driving (AD) systems require rigorous validation across extensive scenarios to ensure robustness. However, real-world data collection is inherently biased towards nominal driving conditions, while safety-critical cases remain rare and expensive to capture at scale. Editing existing driving logs emerges as a scalable alternative, synthesizing diverse training samples while maintaining high fidelity. Simultaneous control over both local object editing and global ego-trajectory editing offers comprehensive flexibility for simulation, as it integrates these distinct edits into a cohesive system, thereby generating reactive scenarios.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

However, unifying these tasks necessitates reconciling two conflicting requirements: generative flexibility for object editing and physical precision for trajectory control. On the one hand, generating high-fidelity details such as object shadows requires powerful generative priors (Hassan et al., 2025; Wang et al., 2025). However, without strict spatial constraints, this generative freedom compromises 3D alignment, causing geometric drift under viewpoint shifts. On the other hand, ensuring view-consistency across ego-trajectories demands explicit 3D geometry (Chen et al., 2025b; Yan et al., 2024). Conversely, lacking generative capacity, these rigid representations cannot plausibly fill occlusions or match lighting, leading to artifacts during object editing. While recent works (Zhao et al., 2025a; Yan et al., 2025) incorporate diffusion into reconstruction, they restrict it to texture refinement rather than semantic creation. Thus, reconciling the structural rigor of trajectory control with the semantic flexibility of object editing remains challenging.

To address this, we introduce SceneDirector, a framework that bridges explicit geometry and generative priors. Specifically, to establish explicit geometry, we construct a Unified Geometric Scaffold by fusing LiDAR-guided depth with editable 3D assets. Eliminating the need for per-scene training, this explicit structure is computationally efficient and guarantees structural fidelity while naturally accommodating 3D asset integration, thereby serving as a unified guide for both object and trajectory control. To leverage generative priors, we formulate the source video as a Static Texture Bank, encoding it into a static key-value memory to provide rich appearance context. Finally, we propose Mask-Gated Reference Attention to bridge these paradigms. It leverages geometric features from the scaffold as spatial queries to attend to the Texture Bank, while utilizing an uncertainty metric to discern reliable geometry from unavoidable occlusions in the scaffold. This gating preserves the reliable structural layout, while dynamically regulating the injection to enhance photorealism in semantically complex regions.

SceneDirector enables simultaneous object editing (insertion, deletion, replacement, repositioning) and ego-trajectory editing within a single inference pass. For object editing, it is agnostic to asset provenance, integrating inputs ranging from scanned datasets (Du et al., 2025) to generative assets synthesized via text-to-3D or single-image reconstruction pipelines (e.g., Wu et al., 2025; Xiang et al., 2025). For ego-trajectory editing, it enables free-form viewpoint control, harmonizing geometric precision with generative fidelity. Our main contributions are as follows:

- We propose SceneDirector, a framework unifying editing of objects and the ego-trajectory. By bridging explicit geometry and generative priors, it reconciles structural consistency for trajectory control and photorealistic synthesis for object manipulation.
- We achieve this bridge via a Unified Geometric Scaffold for structural guidance and a Static Texture Bank for appearance context, integrated by Mask-Gated Reference Attention that preserves reliable structure while synthesizing details in semantically complex regions.
- We demonstrate SceneDirector's superior controllability and visual fidelity through extensive experiments.

2. Related Works

2.1. Driving Scene Generation

The evolution of driving scene generation has shifted from structurally conditioned synthesis to predictive world modeling with granular control. Early works (Li et al., 2025a; Gao et al., 2024b;a; Russell et al., 2025; Zhao et al., 2025b) predominantly focused on ensuring spatial consistency by conditioning on explicit 3D priors, such as semantic occupancy grids, HD maps, and bounding boxes. Building upon this, the focus shifted towards predictive world models (Wang et al., 2024b; Zheng et al., 2024; Gao et al., 2024c) and foundation models (Ren et al., 2025) designed to forecast future states with prioritized temporal coherence and video fidelity. To further broaden the functional scope, Orbis (Mousakhan et al., 2025) and Vista (Gao et al., 2024c) address the stability of long-horizon dynamics, whereas DriVerse (Li et al., 2025c) and GEM (Hassan et al., 2025) propose flexible conditioning paradigms leveraging multimodal trajectory prompts and object-centric visual embeddings to facilitate fine-grained manipulation of scene evolution.

2.2. Driving Scene Editing

Methodologies for driving scene editing are generally categorized into local object manipulation and global trajectory synthesis. Composition-based approaches (Chen et al., 2021; Li et al., 2022; Bai et al., 2024) represent the early paradigm, inserting objects via explicit geometric rendering, though often struggling with photorealistic composition. The advent of diffusion models has shifted the focus to generative object editing, where works like DriveEditor (Liang et al., 2025b), GenMM (Singh et al., 2024), and SceneCrafter (Zhu et al., 2025) leverage 3D layout conditions for spatially grounded synthesis. To further enhance geometric fidelity, G^2Editor (Li et al., 2025b) and R3D2 (Ljungbergh et al., 2025) integrate reusable 3D assets or relightable priors into the diffusion process. In parallel, trajectory-centric methods aim to tackle consistency under ego-motion changes. GeoDrive (Chen et al., 2025a) and FreeVS (Wang et al., 2025) utilize geometry-aware warping for novel view synthesis, while DiST-4D (Guo et al., 2025) and Stag-1 (Wang et al., 2024a) explore disentangled spatio-temporal representations to maintain video coherence. However, a unified framework capable of simultaneously

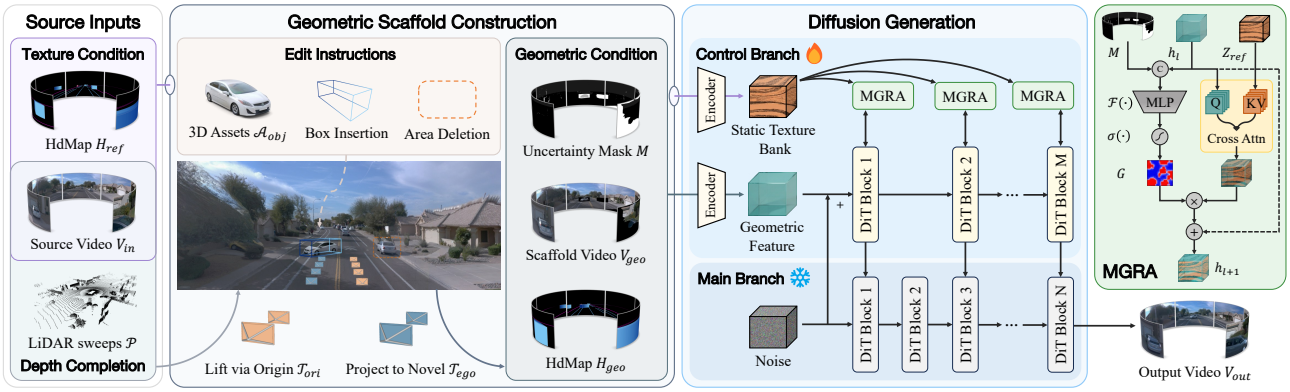

Figure 2. Overview of SceneDirector. **(Left)** Unified Geometric Scaffold Construction: The scaffold is constructed by fusing LiDAR-guided depth with editable 3D assets. It is then rendered under the target ego-trajectory $\mathcal{T}_{ego}$ to yield the Geometric Video V_{geo} (structural guide) and Uncertainty Mask M (reliability metric). **(Right)** Diffusion Generation: The model synthesizes the output video V_{out} using a dual-branch DiT. The texture condition is encoded into a Static Texture Bank to provide appearance context. We introduce the Mask-Gated Reference Attention (MGRA) to bridge explicit geometry and generative priors. It utilizes M to dynamically regulate the interaction between the scaffold and the bank—preserving reliable geometry while adaptively injecting textures for semantic refinement.

handling decoupled object editing and trajectory editing within a single inference pass remains an open challenge.

2.3. Driving Scene Reconstruction

Reconstruction pipelines (Yang et al., 2023; Chen et al., 2025b; Yan et al., 2024; Huang et al., 2024a) decompose scenes into static backgrounds and dynamic actors using neural representations (e.g., NeRF (Mildenhall et al., 2021), 3DGS (Kerbl et al., 2023)) for replay. Some artworks (Wei et al., 2024; Xiong et al., 2025; Lu et al., 2025) provide limited scene editing capabilities but lack visual authenticity. However, these explicit methods are fundamentally limited by input coverage, lacking the generative capacity to plausibly synthesize occluded regions exposed by static object removal or significant trajectory deviations. To address rendering artifacts in novel views, recent hybrid frameworks (Zhao et al., 2025a; Yan et al., 2025; Ni et al., 2025; Mao et al., 2025; Lin et al., 2025; Chen & Peng, 2025) augment 3D representations with video diffusion priors. Crucially, these generative components are primarily optimized for visual enhancement rather than semantic manipulation, focusing on texture refinement rather than inferring content for complex counterfactual edits. In contrast, our approach utilizes a training-free, efficient geometric scaffold to actively guide the generative model, enabling it to synthesize coherent details for substantial scene alterations.

3. Method

Formally, SceneDirector models the conditional generation of a driving video V_{out} given a source video V_{in}, LiDAR sequence $\mathcal{P} = \{P_t\}_{t=1}^{T}$, and a set of user-defined edits. Specifically, object manipulation is defined by target 3D bounding boxes $\mathcal{B}_{obj} \in \mathbb{R}^{T\times N\times 8\times 3}$ with associated 3D assets $\mathcal{A}_{obj}$, while trajectory control is governed by a target ego-pose sequence $\mathcal{T}_{ego} \in \mathbb{R}^{T\times 4\times 4}$. Sec. 3.1 first outlines the Diffusion Transformer backbone. Sec. 3.2 details the training data construction pipeline. Sec. 3.3 details the construction of the Unified Geometric Scaffold, which integrates depth-completed point clouds and 3D assets into a cohesive spatial representation to enforce structural layout. Finally, Sec. 3.4 introduces our Mask-Gated Reference Attention, which preserves reliable structure while dynamically regulating texture injection to synthesize details.

3.1. Preliminaries

Our work builds upon Cosmos-Transfer2.5 (Ali et al., 2025), a state-of-the-art video world model based on Diffusion Transformers (DiT; Peebles & Xie, 2023). Input videos are first compressed into continuous latents x using the Causal 3D VAE in WAN2.1 (Wan et al., 2025). Unlike standard diffusion, the model adopts a Rectified Flow (Liu et al., 2023) objective, defining the forward process as a linear interpolation $x_t = (1-t)x + t\epsilon$ between data x and noise $\epsilon \sim \mathcal{N}(0, I)$. The denoising network v_θ learns to predict the constant velocity field $u_t = \epsilon - x$ by minimizing:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x,\epsilon,\mathbf{c}}\left[\|v_\theta(x_t, t, \mathbf{c}) - (\epsilon - x)\|^2\right], \quad (1)$$

where $\mathbf{c}$ encompasses the set of conditioning signals. Specifically, the model is conditioned on text descriptions c_t encoded by a Vision-Language Model (Azzolini et al., 2025) and spatial controls c_s. We extend the architecture to bridge explicit geometric guidance and generative priors via our proposed Mask-Gated Reference Attention mechanism.

3.2. Self-Supervised Pair Construction

A core challenge in training video editing models is the absence of paired data (i.e., ground truth videos corresponding

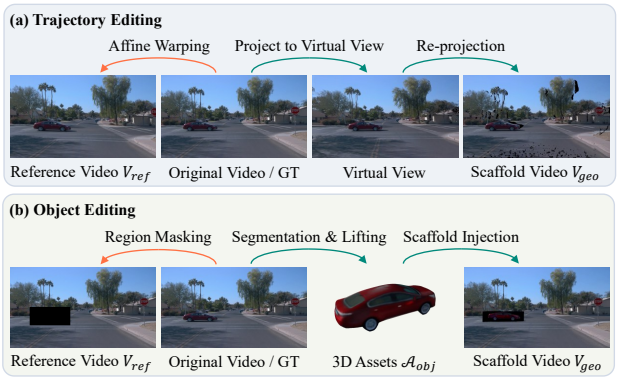

Figure 3. Pipeline for self-supervised pair construction. **(a) Trajectory:** We synthesize a misaligned V_{ref} via affine warping, and generate V_{geo} via round-trip projection to mimic sparsity artifacts. **(b) Object:** We mask targets in V_{ref} to prevent shortcut learning, while composing lifted 3D assets into V_{geo} for structural guidance.

to novel trajectories or modified objects). To address this, we formulate a self-supervised reconstruction task. We treat the original video clip V_{gt} as the reconstruction target and synthesize a training triplets to simulate the geometric and appearance discrepancies encountered during inference.

Synthetic View Perturbation. To mimic the spatial misalignment inherent in trajectory editing, we synthesize a structurally misaligned reference video V_{ref} and a degraded yet aligned scaffold video V_{geo}. As shown in Figure 3 (a), first, we define a perturbation function Φ (detailed in the Appendix D) that covers forward, backward, left, and right movements. These deviations are synthesized via temporal resampling and affine warping, yielding the misaligned $V_{ref} = \Phi(V_{gt})$. Second, to generate aligned V_{geo}, we subject the geometric scaffold to a round-trip projection: the point cloud is first rendered according to Φ to yield intermediate RGB and depth maps. These maps are re-lifted into 3D space and reprojected to the recorded ego-pose. Finally, we rasterize the projected points to generate the scaffold video V_{geo} and the Uncertainty Mask $\mathbf{M}$.

3D Asset Curation. To facilitate object-centric editing, we build a library of high-fidelity 3D assets. We construct an automated filtering pipeline (detailed in the Appendix D) to select objects with sufficient visibility and temporal stability. Valid observations are segmented via SAM2 (Ravi et al., 2025), verified for semantic integrity using Qwen2.5-VL (Bai et al., 2025), and lifted into 3D assets $\mathcal{A}_{obj}$ using Trellis (Xiang et al., 2025). To prevent shortcut learning, we mask the projected regions of $\mathcal{B}_{obj}$ in V_{ref}. Once curated, these assets serve as the source material for both the training and inference pipelines.

Training Pairs Formulation. We formulate the training set as a collection of samples $\mathcal{D} = \{(\mathcal{S}^{(i)}, \mathcal{C}^{(i)}, V_{gt}^{(i)})\}_{i=1}^{N_{data}}$. Specifically, $\mathcal{S} = (V_{geo}, \mathbf{M}, H_{geo})$ denotes the geometric condition, serving as the spatially-aligned structural anchor, while $\mathcal{C} = (V_{ref}, H_{ref})$ represents the texture condition, encapsulating the misaligned appearance cues. H_{geo} and H_{ref} denote the HDMaps corresponding to the recorded and perturbed trajectories, respectively. The model is optimized to reconstruct the ground truth V_{gt} by querying texture from $\mathcal{C}$ and warping it to adhere to the layout dictated by $\mathcal{S}$.

3.3. Unified Geometric Scaffold Construction

To enable precise manipulation of the ego-trajectory and specified target objects, we explicitly construct the Unified Geometric Scaffold—defined as a 3D composition fusing the preserved scene context with manipulable object assets $\mathcal{A}_{obj}$, and render it under the target ego-trajectory $\mathcal{T}_{ego}$ to yield a coherent structural guidance sequence.

Scene Composition and Editing. We formulate the scene at each time step t as a composite of the context $\mathcal{P}_{ctx}$ (encompassing the static background and non-edited entities) and the set of target 3D assets $\mathcal{A}_{obj}$. We build dense scene geometry via a multi-modal depth completion network (DMD^{3}C; Liang et al., 2025a), fusing the RGB image I_t with raw LiDAR sweeps to predict a depth map D_t. These 2D observations are then lifted into the global 3D space. Formally, a pixel $\mathbf{u} = (u, v)$ with predicted depth $d = D_t(\mathbf{u})$ is back-projected to its world coordinate $\mathbf{x}$ via:

$$\mathbf{x} = \mathbf{R}_t \cdot \left(d \cdot \mathbf{K}_{cam}^{-1}[u, v, 1]^\top\right) + \mathbf{t}_t, \tag{2}$$

where $\mathbf{K}_{cam}$ is the camera intrinsic matrix. $\mathbf{T}_t = \{\mathbf{R}_t, \mathbf{t}_t\}$ denotes the ego-pose, which transforms points from the local camera frame to the global world frame.

Projective Rendering under Novel Trajectories. Given a novel ego-trajectory $\mathcal{T}_{ego} = \{\mathbf{T}'_t\}_{t=1}^{T}$ where each $\mathbf{T}'_t \in SE(3)$ represents the target ego-pose, we render point cloud $\mathcal{P}_t$ into a sequence of geometric maps. For an arbitrary point $\mathbf{x}_i \in \mathcal{P}_t$, its projection onto the image plane coordinate $\mathbf{u}_{i,t} = [u, v]^\top$ is governed by the pinhole camera model:

$$s \cdot [u, v, 1]^\top = \mathbf{K}_{cam} \cdot [\mathbf{I}|\mathbf{0}] \cdot (\mathbf{T}'_t)^{-1} \cdot [\mathbf{x}_i^\top, 1]^\top, \tag{3}$$

where s is the scale factor representing depth. We rasterize the composite point cloud into the scaffold video V_{geo}, employing a Z-buffer mechanism to resolve visibility conflicts by retaining the color of the nearest surface at each pixel.

Sparsity and Uncertainty Modeling. Rendering from novel trajectories inevitably introduces disocclusion holes where geometry is completely undefined. Conversely, regions beyond the sensor's field-of-view (e.g., sky) rely entirely on depth completion. While these filled regions lack the precision of raw measurements, they still provide essential low-frequency structural cues. To explicitly quantify the reliability of these geometric cues, we assign a discrete uncertainty category index $\mathbf{M}_t(\mathbf{u}) \in \{0, 1, 2, 3\}$ based on the source of the projected pixel. Let $\mathbf{x}_\mathbf{u}$ denote the nearest

3D surface point visible at pixel $\mathbf{u}$:

$$\mathbf{M}_t(\mathbf{u}) = \begin{cases} 0 & \text{if } \mathbf{x_u} \in \mathcal{P} \quad \text{(Sensor-Verified)} \\ 1 & \text{if } \mathbf{x_u} \in \mathcal{P}_{ctx} \setminus \mathcal{P} \quad \text{(Inferred Layout)} \\ 2 & \text{if } \mathbf{x_u} \in \mathcal{A}_{obj} \quad \text{(Synthetic Asset)} \\ 3 & \text{otherwise} \quad \text{(Voids)} \end{cases} \tag{4}$$

By distinguishing precise sensor measurements from inferred structures, the mask explicitly categorizes the reliability of the geometric confidence. This guidance serves as the core condition for our Mask-Gated Reference Attention.

3.4. Mask-Gated Reference Attention

The Unified Geometric Scaffold enforces rigorous spatial alignment but lacks texture in sensor-denied regions. Conversely, the source video provides rich details yet suffers from spatial misalignment. To bridge this gap, we introduce Mask-Gated Reference Attention (MGRA), an uncertainty-aware mechanism that dynamically regulates texture injection based on local geometric reliability.

We instantiate the Static Neural Texture Bank by encoding the texture condition $\mathcal{C} = (V_{ref}, H_{ref})$ into a frozen latent representation $\mathbf{Z}_{ref}$. To ensure computational efficiency, this representation is shared across all blocks, serving as a static memory queried by the geometric features derived from $\mathcal{S} = (V_{geo}, \mathbf{M}, H_{geo})$. Formally, in each control block l, the intermediate geometric features $\mathbf{h}_l$ serve as the query $\mathbf{Q}_l = \mathbf{h}_l \mathbf{W}_Q$. This query attends to the keys and values $\mathbf{K}, \mathbf{V}$ projected from Z_{ref}, using global context to retrieve corresponding textures despite spatial misalignment.

However, naive attention injection is suboptimal: it risks overwriting precise structural cues in geometrically verified regions while failing to hallucinate details in occluded areas. To address this, we formulate the fusion process as a reliability-aware gating mechanism. We introduce a learned reliability gate $\mathbf{G}$ that modulates the trade-off between the fidelity of the local geometric signal and the richness of the reference bank, explicitly guided by the sensor's epistemic uncertainty. We define this gating function as:

$$\mathbf{G} = \sigma\left(\mathcal{F}\left([\mathbf{h}_l; \Psi(\mathbf{M})]\right)\right), \tag{5}$$

where $[\cdot;\cdot]$ denotes channel concatenation, σ is the Sigmoid activation, and $\mathcal{F}$ is a multilayer perceptron (MLP). $\Psi(\cdot)$ denotes an embedding operator that downsamples $\mathbf{M}$ to match the latent resolution and projects the category indices into a continuous feature space using a learnable embedding layer. Jointly conditioning on $\mathbf{M}$ and $\mathbf{h}_l$ allows the gate to be content-adaptive, enabling the mechanism to discern semantic complexity even where sensor coverage is uniform.

The retrieved texture information is then injected into the control stream via a modulated residual connection. The feature update process is formulated as:

$$\mathbf{h}_{l+1} = \mathbf{h}_l + \lambda \cdot \left(\mathbf{G} \odot \text{Attention}(\mathbf{Q}_l, \mathbf{K}, \mathbf{V})\right), \tag{6}$$

where $\odot$ represents element-wise multiplication and λ is a zero-initialized learnable scaling factor. This mechanism naturally induces a spatially adaptive disentanglement: the gate suppresses the reference stream ($\mathbf{G} \to 0$) in structurally homogeneous regions to strictly enforce geometric constraints, while activating ($\mathbf{G} \to 1$) in semantically rich regions to facilitate texture synthesis from the bank.

4. Experiments

4.1. Object editing

Evaluation Benchmark. To benchmark capabilities, we curate 64 representative evaluation scenarios from the Waymo Open Dataset (WOD; Sun et al., 2020) validation split via a fully automated pipeline (detailed in Appendix C). We establish three protocols: **Single-Edit** (insertion) isolates geometric precision and fidelity from sequential errors. **Multi-Edit** evaluates the ability to synthesize complex scenarios requiring simultaneous multi-type editing. **Unified Capability Analysis** specifically validates the unified editing performance by executing object manipulation simultaneously with ego-trajectory modification (2m Gradual Transition).

Baselines and Protocol. We compare SceneDirector against two baselines: *VACE-14B* (Jiang et al., 2025), a large-scale general-domain video editor, and *DriveEditor* (Liang et al., 2025b), a domain-specific editor guided by 3D layouts. Since both baselines are limited to single-view, single-object editing, we apply them sequentially (deletion $\to$ replacement $\to$ repositioning $\to$ insertion) for Multi-Edit comparisons, whereas SceneDirector performs unified editing in a single inference pass. As our method inherently requires multi-view inputs, we align the evaluation by comparing our front-view results against the single-view baselines.

Metrics. We evaluate generation quality using FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018), and semantic alignment via CLIP-I (Huang et al., 2024b). To evaluate the downstream perception performance, we employ a PGD detector (Wang et al., 2022) pre-trained on WOD. We report detection Recall to assess the efficacy of the edits, along with Average Translation Error (ATE) and Average Orientation Error (AOE) to quantify geometric fidelity.

Main Results. Table 1 and Figure 4 presents the comparison. (1) **Versatile Editing:** To ensure a fair comparison, we establish a Single-Edit protocol that aligns with the baselines' native capabilities. In *Single-Edit*, SceneDirector ensures superior geometric alignment (ATE 0.78m). While VACE-14B shows a slight advantage in texture metrics (FID) attributed to its massive 14B parameter capacity, it lacks precise geometric controllability. In *Multi-Edit*, our method leads in all metrics (FVD 516.83 vs. 729.42), effectively eliminating sequential error accumulation. (2) **Unified Capability:** We analyze simultaneous object and

Figure 4. Qualitative comparison on multi-object editing. Each row illustrates a scenario combining two distinct editing operations simultaneously. **Left to Right:** Original frames; target 3D assets (crossed-out items denote deletion); and 3D layout conditions (Blue: Insertion, Green: Replacement). **Observations:** *DriveEditor* suffers from residual artifacts (e.g., gray patches) due to ineffective masking strategies. *VACE* struggles with precise geometric alignment and may erroneously hallucinate objects during target deletion. *SceneDirector* synthesizes high-fidelity assets that are strictly aligned with the 3D bounding boxes and harmoniously integrated into the scene context.

Table 1. Quantitative comparison on Object Editing. **Multi-Edit** evaluates the unified framework on complex scenarios. **Single-Edit** isolates intrinsic generation quality, confirming our geometric precision (lowest ATE/AOE). The bottom rows confirm that **simultaneous editing** (*Obj.+Traj.*) maintains high precision comparable to object-only editing (*Obj. Only*), validating our unified capability.

Method	Multi-Edit					Single-Edit				
	FID ↓	FVD ↓	CLIP-I ↑	ATE ↓	AOE ↓	FID ↓	FVD ↓	CLIP-I ↑	ATE ↓	AOE ↓
VACE	49.18	729.42	76.36	1.09	0.077	**35.04**	**451.31**	76.98	1.02	0.075
DriveEditor	51.43	818.48	75.98	0.93	0.074	42.75	556.59	76.14	0.86	0.071
SceneDirector (Ours)	**38.29**	**516.83**	**76.85**	**0.81**	**0.052**	35.83	464.56	**77.02**	**0.78**	**0.052**
Ours (Multi-View, Obj. Only)	35.18	305.11	76.78	0.90	0.055			–		
Ours (Multi-View, Obj. + Traj.)	36.65	326.16	76.40	0.95	0.056			–		

trajectory editing (2m Gradual Transition). Compared to the static-trajectory multi-view baseline (*Obj. Only*), the unified setting (*Obj. + Traj.*) exhibits negligible degradation (ATE 0.90m → 0.95m). This variance stems from detection instability under moving viewpoints rather than generative failure, confirming that our scaffold effectively decouples tasks, ensuring robust unified editing without interference.

4.2. Ego-trajectory Editing

Evaluation Benchmark. To assess ego-trajectory controllability, we curate 64 evaluation scenarios from the WOD validation split via an automated pipeline (detailed in the Appendix C). Each scenario is evaluated under two deviation magnitudes (2m and 3m) and two trajectory modification modes: *Gradual Transition*, which simulates a gradual deviation from the original path (e.g., lane changing), and *Fixed Offset*, which maintains a constant spatial displacement.

Baselines. We evaluate ego-trajectory editing against five baselines across two categories. **(1) Reconstruction-based frameworks** include *StreetGaussian* (Yan et al., 2024), which utilizes dynamic 3DGS, and *StreetCrafter* (Yan et al., 2025), a hybrid method combining diffusion priors with 3DGS. To ensure a rigorous comparison against both reconstruction and hybrid baselines, we introduce *SceneDirector+SG*, training StreetGaussian on our videos under standard StreetCrafter settings. This serves two aims: (i) against StreetGaussian, it evaluates if our videos provide effective supervision to enhance reconstruction; and (ii) against StreetCrafter, it aligns with the hybrid paradigm, isolating the generative quality to benchmark the diffusion priors. **(2) Diffusion-based methods** include *FreeVS* (Wang et al., 2025), using pseudo-view priors for guidance; *GEM* (Hassan et al., 2025), a trajectory-guided world model; and *StreetCrafter-DM*, the standalone diffusion component of StreetCrafter. Note that GEM and StreetCrafter-DM are excluded from settings incompatible with their control modes.

Metrics. We assess performance across generation quality and trajectory controllability. For visual quality, we likewise report FID and FVD. To quantify structural preservation under trajectory manipulation, we utilize a pre-trained 3D lane detector, Persformer (Chen et al., 2022), on the front-view videos. We treat lane detections from the source video as ground truth, which ensures a consistent relative benchmark because all methods are evaluated under identical deviation settings. For generated videos, the predicted lane points are projected back to the source coordinate system to measure alignment. We report F1-score, Recall, and X-error (lateral error in meters) to evaluate the geometric accuracy.

Main Results. Tables 2, 3 and Figure 5 present trajectory editing performance. SceneDirector outperforms diffusion-based baselines, showing superior visual fidelity (FID 34.48, FVD 476.1) compared to FreeVS (FVD 1208.9) and GEM (FVD 667.9). Using the geometric scaffold ensures rig-

Figure 5. Qualitative comparison of trajectory editing under **Gradual Transition**. *StreetGaussian* fails to generalize to novel poses, showing severe distortions and artifacts. While *StreetCrafter* improves texture via diffusion (*StreetCrafter-DM*), it lacks explicit in-painting capabilities for disoccluded regions (e.g., behind trees), resulting in gray voids due to missing geometry. *GEM* displays weak control fidelity with noticeable structural loss. *FreeVS* produces incomplete frames by masking out the sky due to sparse LiDAR coverage. *SceneDirector* effectively hallucinates plausible details in geometric voids (e.g., sky and occluded background) while maintaining rigorous structural consistency, yielding artifact-free video synthesis. Additional results for **Fixed Offset** are provided in the Appendix E.

Table 2. Visual quality of trajectory editing. In the front-view setting, SceneDirector consistently outperforms all diffusion-based baselines. SceneDirector+SG performs best under 2m gradual deviation. SceneDirector-MV surpasses all competing methods in the multi-view.

Method	Category	Gradual Transition				Fixed Offset			
		2m Deviation		3m Deviation		2m Offset		3m Offset	
		FID ↓	FVD ↓	FID ↓	FVD ↓	FID ↓	FVD ↓	FID ↓	FVD ↓
FreeVS	Diffusion	79.25	1208.9	79.52	1255.5	81.34	1229.9	84.65	1214.2
GEM		42.90	667.9	44.02	661.2	-	-	-	-
StreetCrafter-DM		-	-	-	-	46.25	694.0	52.22	757.7
SceneDirector (Ours)		34.48	476.1	37.21	513.5	36.38	477.3	46.96	564.9
StreetGaussian	Reconstruction	27.37	434.9	35.64	535.0	43.02	585.2	59.97	774.4
StreetCrafter		26.36	418.2	32.44	468.4	**33.32**	**417.7**	**43.27**	**539.9**
SceneDirector+SG (Ours)		**25.70**	**400.1**	**31.24**	**453.5**	35.64	449.5	45.33	566.0
FreeVS-MV	Multiview	69.05	898.5	69.43	927.3	73.83	936.5	75.06	940.5
StreetGaussian-MV		54.22	549.5	58.86	643.2	68.25	701.3	78.15	868.1
SceneDirector-MV (Ours)		**44.41**	**445.5**	**49.10**	**469.2**	**51.34**	**474.1**	**57.05**	**544.6**

Table 3. Geometric accuracy of front-view trajectory editing (X-ERR in meters). SceneDirector outperforms diffusion-based methods in structural alignment, while SceneDirector+SG exceeds reconstruction baselines in lane alignment, confirming high structural accuracy.

Method	Gradual Transition						Fixed Offset					
	2m Deviation			3m Deviation			2m Offset			3m Offset		
	F1 ↑	R ↑	X-err ↓	F1 ↑	R ↑	X-err ↓	F1 ↑	R ↑	X-err ↓	F1 ↑	R ↑	X-err ↓
FreeVS	40.3	38.7	0.814	37.0	34.2	0.867	42.6	36.0	0.712	44.9	43.7	0.817
GEM	19.4	17.6	1.130	17.7	15.8	1.129	-	-	-	-	-	-
StreetCrafter-DM	-	-	-	-	-	-	40.8	33.7	0.860	35.5	27.9	0.902
SceneDirector (Ours)	55.1	52.5	0.603	51.4	46.3	0.640	54.0	50.2	0.659	46.6	42.7	0.768
StreetGaussian	59.2	54.0	0.535	53.1	45.9	**0.576**	55.4	48.1	**0.580**	42.0	33.3	**0.722**
StreetCrafter	59.5	54.9	0.541	54.1	48.4	0.596	56.0	51.0	0.631	46.8	**41.3**	0.744
SceneDirector+SG (Ours)	**60.4**	**56.7**	**0.528**	**55.4**	**50.4**	0.590	**57.4**	**53.6**	0.622	**47.5**	41.2	0.740

orous structural alignment (F1 55.1), surpassing FreeVS (F1 40.3). SceneDirector+SG outperforms StreetCrafter in Gradual Transition (FID 25.70 vs. 26.36). While maintaining comparable visual quality in Fixed Offset, it achieves superior structural alignment across all settings (e.g., F1 47.5 vs. 46.8 at 3m). This superior reconstruction validates our multi-view consistency and geometric correctness under viewpoint shifts, as 3DGS convergence demands strict geometric consensus. This shows that SceneDirector not only excels as a standalone diffusion model, but also provides high-quality priors that enhance reconstruction. Please refer to Appendix B for evaluation of multi-view consistency.

4.3. Ablation Study

We validate the components on the multi-view trajectory editing task (2m fixed offset and 3m gradual transition) using FID and FVD. We select this task as global view synthesis is more discriminative than localized object editing.

Impact of Reference Injection. As shown in Table 4, removing the Mask-Gated Reference Attention (MGRA) (*w/o Ref. Attn*) leads to a performance drop, with FVD increasing

Figure 6. Qualitative results of SceneDirector on multi-view editing. **Top (Object Editing):** Rows display inputs (object assets, 3D bounding boxes), source frames, and our editing results. Our method enables unified editing in a single inference pass. **Bottom (Trajectory Editing):** Results under a gradual right shift demonstrate that our method preserves cross-view structural consistency and photorealism.

Table 4. Ablation study. **w/o Ref. Attn**: Removing the reference attention. **Ungated**: Using standard cross-attention without the uncertainty mask. **Alt. Geo**: Using an alternative depth completion.

Method	3m Deviation		2m Offset	
	FID ↓	FVD ↓	FID ↓	FVD ↓
SceneDirector	**49.10**	**469.2**	**51.34**	**474.1**
1. w/o Ref. Attn	53.02	495.0	55.46	512.5
2. Standard Attn	50.85	481.8	53.02	493.4
3. Alt. Geometry	50.04	477.7	51.87	478.0

by +25.8 (3m Deviation). This confirms that appearance priors are indispensable for visual quality, particularly to compensate for information loss in sparse point clouds.

Necessity of Uncertainty-Aware Gating. As shown in Figure 7, replacing MGRA with standard cross-attention (*Standard Attn*) results in severe ghosting artifacts. Visualization of the gate maps reveals a noise-dependent adaptive mechanism. At early timesteps, high noise levels dominate h_l, forcing the gate to rely primarily on M. As denoising resolves semantic structures, the gate is content-adaptive: it discerns semantic complexity, selectively injecting texture details into rich areas (e.g., vehicles, buildings), rather than strictly adhering to the geometric confidence.

Robustness to Geometric Quality. Finally, replacing the depth completion backbone with LRRU (Wang et al., 2023) (*Alt. Geo*) yields only marginal metric degradation (FID 51.34 → 51.87 in 2m Fixed Offset). While the quality drop is more pronounced under the challenging 3m Deviation setting (FID 49.10 → 50.04), this stability confirms that SceneDirector accommodates scaffolds of diverse qualities.

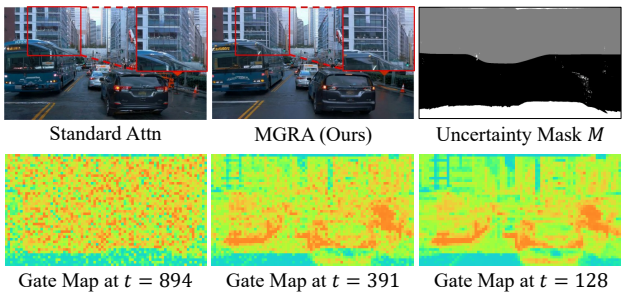

Figure 7. Ablation study on Mask-Gated Reference Attention (MGRA). **(Top)** MGRA eliminates ghosting artifacts observed in standard cross-attention. We visualize the mask M (Black: sensor-verified; Gray: inferred layout; White: voids). **(Bottom)** Evolution of gate values across diffusion timesteps t, where red indicates higher gate activation. The gating mechanism shifts from reliance on the mask prior under high noise ($t = 894$) to content adaptation ($t = 128$), selectively retrieving textures for semantic regions.

5. Conclusion

We propose SceneDirector, a unified framework that bridges explicit geometry and generative priors. This reconciles the structural consistency required for trajectory control with the photorealistic synthesis essential for object manipulation. Key to this success is Mask-Gated Reference Attention, which leverages sensor uncertainty to harmonize the Unified Geometric Scaffold with generative texture hallucination. Extensive evaluations on the Waymo Open Dataset demonstrate the superior controllability and quality of SceneDirector, offering a scalable solution for simultaneous objects and ego-trajectory editing. We provide a detailed discussion on limitations and failure cases in Appendix G.

Impact Statement

This paper presents SceneDirector, a framework designed to advance the validation and robustness of autonomous driving systems by synthesizing diverse and safety-critical driving scenarios. By enabling the generation of rare "corner cases" without extensive physical testing, our work has the potential to improve road safety and reduce the environmental costs associated with real-world data collection.

However, we acknowledge that the generative capabilities of our method—specifically the ability to realistically alter vehicle trajectories and manipulate scene objects—carry potential risks. Like other high-fidelity video editing technologies, this framework could be misused to create misleading media or fabricate evidence (e.g., altered dashcam footage). Furthermore, relying on synthetic data for safety-critical applications requires rigorous validation to ensure that generated scenarios maintain physical realism and do not introduce biases that could compromise system reliability in the real world. We encourage the research community to develop robust detection mechanisms for synthetic media and to maintain strict oversight when utilizing generated data for safety validation.

References

Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.-W., et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Bai, C., Shao, Z., Zhang, G., Liang, D., Yang, J., Zhang, Z., Guo, Y., Zhong, C., Qiu, Y., Wang, Z., et al. Anything in any scene: Photorealistic video object insertion. *arXiv preprint arXiv:2401.17509*, 2024.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Chen, A., Zheng, W., Wang, Y., Zhang, X., Zhan, K., Jia, P., Keutzer, K., and Zhang, S. Geodrive: 3d geometry-informed driving world model with precise action control. *arXiv preprint arXiv:2505.22421*, 2025a.

Chen, L., Sima, C., Li, Y., Zheng, Z., Xu, J., Geng, X., Li, H., He, C., Shi, J., Qiao, Y., et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision (ECCV)*, pp. 550–567. Springer, 2022.

Chen, S. and Peng, P. Freegen: Feed-forward reconstruction-generation co-training for free-viewpoint driving scene synthesis. *arXiv preprint arXiv:2512.04830*, 2025.

Chen, Y., Rong, F., Duggal, S., Wang, S., Yan, X., Manivasagam, S., Xue, S., Yumer, E., and Urtasun, R. Geosim: Realistic video simulation via geometry-aware composition for self-driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7230–7240, 2021.

Chen, Y., Gu, C., Jiang, J., Zhu, X., and Zhang, L. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv preprint arXiv:2311.18561*, 2025b.

Du, X., Wang, Y., Sun, H., Wu, Z., Sheng, H., Wang, S., Ying, J., Lu, M., Zhu, T., Zhan, K., and Yu, X. 3drealcar: An in-the-wild rgb-d car dataset with 360-degree views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26488–26498, October 2025.

Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., and Xu, Q. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024a.

Gao, R., Chen, K., Xie, E., HONG, L., Li, Z., Yeung, D.-Y., and Xu, Q. Magicdrive: Street view generation with diverse 3d geometry control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024b.

Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., and Li, H. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 91560–91596, 2024c.

Guo, J., Ding, Y., Chen, X., Chen, S., Li, B., Zou, Y., Lyu, X., Tan, F., Qi, X., Li, Z., et al. Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. *arXiv preprint arXiv:2503.15208*, 2025.

Hassan, M., Stapf, S., Rahimi, A., Rezende, P. M. B., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., Cannici, M., Aljalbout, E., Ye, B., Wang, X., Davtyan, A., Salzmann, M., Scaramuzza, D., Pollefeys, M., Favaro, P., and Alahi, A. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22404–22415, June 2025.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Huang, N., Wei, X., Zheng, W., An, P., Lu, M., Zhan, W., Tomizuka, M., Keutzer, K., and Zhang, S. S^3gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024a.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., and Liu, Z. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, June 2024b.

Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., and Liu, Y. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):139–1, 2023.

Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., Zhou, S., Zhang, L., Qi, X., Zhao, H., Yang, M., Zeng, W., and Jin, X. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11971–11981, June 2025a.

Li, J., Jiang, J., Miao, J., Long, M., Wen, T., Jia, P., Liu, S., Yu, C., Liu, M., Cai, Y., et al. Realistic and controllable 3d gaussian-guided object editing for driving video generation. *arXiv preprint arXiv:2508.20471*, 2025b.

Li, N., Song, F., Zhang, Y., Liang, P., and Cheng, E. Traffic context aware data augmentation for rare object detection in autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4548–4554. IEEE, 2022.

Li, X., Wu, C., Yang, Z., Xu, Z., Liang, D., Zhang, Y., Wan, J., and Wang, J. Driverse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment. *arXiv preprint arXiv:2504.18576*, 2025c.

Liang, Y., Hu, Y., Shao, W., and Fu, Y. Distilling monocular foundation model for fine-grained depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22254–22265, June 2025a.

Liang, Y., Yan, Z., Chen, L., Zhou, J., Yan, L., Zhong, S., and Zou, X. Driveeditor: A unified 3d information-guided framework for controllable object editing in driving scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 5164–5172, 2025b.

Lin, J., Wang, K., Wang, S., Fan, S., Li, G., and Gao, W. Vgd: Visual geometry gaussian splatting for feed-forward surround-view driving reconstruction. *arXiv preprint arXiv:2510.19578*, 2025.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Ljungbergh, W., Taveira, B., Zheng, W., Tonderski, A., Peng, C., Kahl, F., Petersson, C., Felsberg, M., Keutzer, K., Tomizuka, M., et al. R3d2: Realistic 3d asset insertion via diffusion for autonomous driving simulation. *arXiv preprint arXiv:2506.07826*, 2025.

Lu, S., Lin, Z., Lu, C., Wang, H., Zhuo, G., and Zheng, L. Multieditor: Controllable multimodal object editing for driving scenarios using 3d gaussian splatting priors. *arXiv preprint arXiv:2507.21872*, 2025.

Mao, J., Li, B., Ivanovic, B., Chen, Y., Wang, Y., You, Y., Xiao, C., Xu, D., Pavone, M., and Wang, Y. Dreamdrive: Generative 4d scene modeling from street view images. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 367–374, 2025.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Mousakhan, A., Mittal, S., Galesso, S., Farid, K., and Brox, T. Orbis: Overcoming challenges of long-horizon prediction in driving world models. *arXiv preprint arXiv:2507.13162*, 2025.

Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Huang, G., Liu, C., Chen, Y., Wang, Y., Zhang, X., Zhan, Y., Zhan, K., Jia, P., Lang, X., Wang, X., and Mei, W. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1559–1569, June 2025.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollar, P., and Feichtenhofer, C. SAM 2: Segment anything in images and videos. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Ren, X., Lu, Y., Cao, T., Gao, R., Huang, S., Sabour, A., Shen, T., Pfaff, T., Wu, J. Z., Chen, R., et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., and Corrado, G. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Singh, B., Kulharia, V., Yang, L., Ravichandran, A., Tyagi, A., and Shrivastava, A. Genmm: Geometrically and temporally consistent multimodal data generation for video and lidar. *arXiv preprint arXiv:2406.10722*, 2024.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, L., Zheng, W., Du, D., Zhang, Y., Ren, Y., Jiang, H., Cui, Z., Yu, H., Zhou, J., Lu, J., et al. Stag-1: Towards realistic 4d driving simulation with video generation model. *arXiv preprint arXiv:2412.05280*, 2024a.

Wang, Q., Fan, L., Wang, Y., Chen, Y., and Zhang, Z. FreeVS: Generative view synthesis on free driving trajectory. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Wang, T., ZHU, X., Pang, J., and Lin, D. Probabilistic and geometric depth: Detecting objects in perspective. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 164, pp. 1475–1485. PMLR, 08–11 Nov 2022.

Wang, Y., Li, B., Zhang, G., Liu, Q., Gao, T., and Dai, Y. Lrru: Long-short range recurrent updating networks for depth completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9422–9432, October 2023.

Wang, Y., He, J., Fan, L., Li, H., Chen, Y., and Zhang, Z. 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 (CVPR)*, pp. 14749–14759, June 2024b.

Wei, Y., Wang, Z., Lu, Y., Xu, C., Liu, C., Zhao, H., Chen, S., and Wang, Y. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15077–15087, June 2024.

Wu, T., Zheng, C., Guan, F., Vedaldi, A., and Cham, T.-J. Amodal3r: Amodal 3d reconstruction from occluded 2d images. *arXiv preprint arXiv:2503.13439*, 2025.

Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., and Yang, J. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, June 2025.

Xiong, Y., Zhou, X., Wan, Y., Sun, D., and Yang, M.-H. Drivinggaussian++: Towards realistic reconstruction and editable simulation for surrounding dynamic driving scenes, 2025. URL https://arxiv.org/abs/2508.20965.

Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., and Peng, S. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision (ECCV)*, pp. 156–173, 2024.

Yan, Y., Xu, Z., Lin, H., Jin, H., Guo, H., Wang, Y., Zhan, K., Lang, X., Bao, H., Zhou, X., and Peng, S. Streetcrafter: Street view synthesis with controllable video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 822–832, June 2025.

Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.-C., Yang, A. J., and Urtasun, R. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1389–1399, June 2023.

Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., Huang, G., Chen, X., Wang, B., Zhang, Y., Mei, W., and

Wang, X. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12015–12026, June 2025a.

Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., and Wang, X. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 10412–10420, 2025b.

Zheng, W., Song, R., Guo, X., Zhang, C., and Chen, L. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision (ECCV)*, pp. 87–104. Springer, 2024.

Zhu, Z., Zou, Y., Jiang, C. M., Sun, B., Casser, V., Huang, X., Wang, J., Yang, Z., Gao, R., Guibas, L., et al. Scenecrafter: Controllable multi-view driving scene editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6812–6822, 2025.

A. Network Architecture and Implementation Details

A.1. Mask-Gated Reference Attention (MGRA)

The MGRA module dynamically modulates texture injection based on geometric uncertainty. The specific hyper-parameters are as follows:

- **Uncertainty Embedding (Ψ):** The discrete uncertainty mask M is projected into a continuous embedding space with a dimension of 256.
- **Gating Network ($\mathcal{F}$):** The MLP accepts a concatenated input of feature and mask embedding. It consists of a linear projection to a hidden dimension of 512, followed by a SiLU activation, and a final projection to a scalar output with Sigmoid activation.
- **Reference Attention:** We employ Multi-Head Cross-Attention with 16 heads. Each head has a dimension of 128 $(2048/16)$.
- **Zero Initialization:** The modulation parameters (scale and shift) for injecting the reference branch are initialized to zero, ensuring the model initially preserves the original geometric structure.

A.2. Training Configuration

We generated 2364 clips for the joint object and trajectory editing task following the preparation strategy in Sec. 3.2. In parallel, we retained 2394 original clips dedicated to the trajectory editing sub-task. We utilize the pre-trained weights of Cosmos-Transfer2.5 with base blocks frozen. The model is trained on 8 NVIDIA A800 GPUs for 40,000 iterations with a learning rate of 2×10^{-5}. We utilize a LambdaLinear scheduler with a linear warmup for the first $1,000$ steps. The effective global batch size is 4. This is achieved using a data parallel size of 4 with context parallel size 2. We use the FusedAdamW optimizer with weight decay $\lambda = 0.1$, and $\epsilon = 10^{-8}$. Training clips are processed at a resolution of 704×1280 with a sequence length of 29 frames.

B. Cross-View Consistency Evaluation

To explicitly quantify the spatial coherence between adjacent camera views, we conduct a direct feature-level comparison in overlapping regions. While the reconstruction metrics in the main paper implicitly validate geometric consensus, this section provides a direct measurement of visual and semantic alignment across sensor boundaries.

B.1. Evaluation Metrics and Setup

Overlap Definition and Camera Setup. The Waymo Open Dataset utilizes a surround-view system consisting of five cameras: Side Left (SL), Front Left (FL), Front (F), Front Right (FR), and Side Right (SR). We evaluate consistency across the four adjacent overlapping interfaces: **SL-FL**, **FL-F**, **F-FR**, and **FR-SR**. For each adjacent pair, we crop the overlapping regions located at their shared boundaries (rightmost 1/3 of the left view and leftmost 1/3 of the right view).

Metrics. We employ two distinct feature extractors:

- **CLIP-I (Semantic Consistency):** Measures the cosine similarity of high-level semantic embeddings. High scores indicate consistent object identities across views.
- **DINOv2 (Structural Consistency):** Measures the cosine similarity of fine-grained geometric features. High scores indicate precise structural alignment without "drifting".

B.2. Quantitative Results

The evaluation results are presented in Table 5 (Semantic) and Table 6 (Structural).

For **Object Editing**, SceneDirector demonstrates high fidelity, achieving consistency scores (e.g., 0.835 CLIP avg) that closely approach the *Original Data* baseline (0.858 avg), verifying that our inserted assets maintain rigorous continuity across views.

For **Trajectory Editing**, we observe distinct behaviors across scenarios:

Table 5. **Semantic Consistency Evaluation (CLIP-I).** We measure the cosine similarity (↑) of CLIP features in overlapping regions. The "Original Data" serves as the real-world reference. *SceneDirector* achieves superior performance in Gradual Transition scenarios. In Fixed Offset settings, while slightly trailing the reconstruction-based StreetGaussian, our method significantly outperforms the diffusion-based FreeVS, demonstrating robust semantic stability against large view shifts.

Task	Scenario	Method	SL-FL	FL-F	F-FR	FR-SR	Avg.
Object Editing	-	Original Data	0.847	0.866	0.867	0.853	0.858
		SceneDirector	**0.838**	**0.861**	**0.863**	**0.839**	**0.850**
Trajectory Editing	Real Data	Original Data	0.863	0.878	0.874	0.859	0.868
	Gradual 2m Dev.	FreeVS-MV	0.857	0.881	0.875	0.844	0.864
		StreetGaussian-MV	0.874	0.874	0.869	0.876	0.873
		SceneDirector	**0.863**	**0.889**	**0.887**	**0.873**	**0.878**
	Gradual 3m Dev.	FreeVS-MV	0.854	0.882	0.872	0.841	0.862
		StreetGaussian-MV	0.872	0.873	0.866	0.876	0.872
		SceneDirector	**0.863**	**0.889**	**0.886**	**0.875**	**0.878**
	Fixed 2m Off.	FreeVS-MV	0.774	0.784	0.782	0.794	0.783
		StreetGaussian-MV	0.861	0.884	0.886	0.873	0.876
		SceneDirector	**0.849**	**0.874**	**0.869**	**0.837**	**0.857**
	Fixed 3m Off.	FreeVS-MV	0.848	0.871	0.868	0.837	0.856
		StreetGaussian-MV	0.865	0.885	0.887	0.872	0.877
		SceneDirector	**0.869**	**0.870**	**0.869**	**0.874**	**0.870**

Table 6. **Structural Consistency Evaluation (DINOv2).** We measure the cosine similarity (↑) of DINOv2 features to assess geometric alignment. The "Original Data" indicates the inherent geometric coherence of the real world. *SceneDirector* maintains structural integrity remarkably close to this real-world baseline across all scenarios. Notably, in the challenging Fixed Offset settings where the diffusion-based FreeVS suffers from severe geometric drift (e.g., dropping to 0.683 avg in Fixed 2m), our method remains robust (0.789 avg), confirming the effectiveness of our Unified Geometric Scaffold.

Task	Scenario	Method	SL-FL	FL-F	F-FR	FR-SR	Avg.
Object Editing	-	Original Data	0.729	0.806	0.806	0.735	0.769
		SceneDirector	**0.727**	**0.819**	**0.822**	**0.727**	**0.774**
Trajectory Editing	Real Data	Original Data	0.765	0.834	0.829	0.763	0.798
	Gradual 2m Dev.	FreeVS-MV	0.722	0.842	0.831	0.704	0.775
		StreetGaussian-MV	0.746	0.832	0.821	0.731	0.783
		SceneDirector	**0.768**	**0.842**	**0.842**	**0.725**	**0.794**
	Gradual 3m Dev.	FreeVS-MV	0.821	0.843	0.831	0.802	0.824
		StreetGaussian-MV	0.744	0.831	0.821	0.727	0.781
		SceneDirector	**0.767**	**0.844**	**0.842**	**0.722**	**0.794**
	Fixed 2m Off.	FreeVS-MV	0.664	0.706	0.703	0.659	0.683
		StreetGaussian-MV	0.741	0.824	0.821	0.728	0.779
		SceneDirector	**0.762**	**0.839**	**0.838**	**0.717**	**0.789**
	Fixed 3m Off.	FreeVS-MV	0.716	0.842	0.833	0.699	0.772
		StreetGaussian-MV	0.737	0.823	0.821	0.724	0.776
		SceneDirector	**0.762**	**0.833**	**0.837**	**0.717**	**0.787**

- **Semantic Stability (Table 5):** Our method achieves superior performance in *Gradual Transition* scenarios. While reconstruction-based methods (StreetGaussian) exhibit strong stability in *Fixed Offsets* due to their explicit representation, SceneDirector remains highly competitive and significantly outperforms FreeVS, ensuring that synthesized visual concepts remain consistent even when unseen regions are hallucinated.

- **Structural Integrity (Table 6):** Pure diffusion baselines (FreeVS) suffer from severe geometric drift, particularly in challenging *Fixed Offset* settings where the average DINOv2 score drops significantly to 0.683. In contrast, SceneDirector maintains robust structural alignment (0.789 avg), tracking the *Original Data* reference (0.798 avg) with remarkable precision. This explicitly validates the effectiveness of our Unified Geometric Scaffold in preventing spatial collapse during large viewpoint shifts.

C. Automated Evaluation Benchmark Construction

C.1. Scenario Curation and Pre-processing

We applied our automated generation pipeline to the validation split. Since not every driving scene physically accommodates all editing types (e.g., due to specific road topology or traffic congestion), the pipeline first filtered for feasible operations. From the pool of successfully generated samples, we curated 64 representative scenarios for object editing and 64 scenarios for trajectory editing. These selected scenarios cover diverse driving conditions, including varying traffic densities, complex turns, and straight roads, ensuring a balanced evaluation.

C.2. Object Editing Pipeline

For each processed clip, the pipeline identifies valid manipulation targets based on temporal stability (tracked ≥ 16 frames) and semantic integrity (Cars/Trucks).

1. Object Insertion. Insertion requires placing new objects that adhere to traffic rules and physical constraints. We employ a multi-stage heuristic planner:

- **Candidate Generation:** We generate candidate trajectories relative to a *Reference Anchor* (the dominant traffic flow or ego-vehicle) using three strategies in descending priority:
  1. *Longitudinal Gap Filling:* The pipeline scans for longitudinal gaps between the reference vehicle and its neighbors. If a gap exceeds the safety threshold ($L_{obj} + 6.0m$), the candidate is placed at the gap's midpoint.
  2. *Platoon Formation:* Candidates are placed at fixed longitudinal offsets (e.g., $\pm 10m$) to simulate car-following.
  3. *Adjacent Lane Injection:* Candidates are spawned in adjacent lanes with lateral offsets of $\pm 3.6m$.
- **Ground Alignment:** We estimate the local ground elevation $z_{ground}(t)$ by interpolating the vertical positions of neighboring vehicles to prevent floating artifacts.
- **Physics and Visibility Validation:** Candidates are accepted only if they pass collision avoidance (OBB intersection test) and frustum visibility checks (visible pixel area > 200 pixels).

2. Object Repositioning. Objects are shifted along their local axes with a fixed magnitude of $\Delta d = 3.0m$. The new position undergoes strict collision checks against static background geometry and dynamic agents.

3. Object Deletion and Replacement. The algorithm automatically identifies prominent foreground objects for **Deletion** (removing rendering instructions) or **Replacement** (swapping 3D asset IDs while maintaining original trajectories), creating pairs for inpainting and semantic generation evaluation.

C.3. Ego-trajectory Editing Pipeline

For ego-trajectory editing, we synthesize novel view sequences under two modes: *Fixed Offset* and *Gradual Transition*. To ensure physical realism, we implement a strict feasibility screening and a kinematic-aware trajectory parameterization.

1. Feasibility Screening via Occupancy Analysis. Before generation, we define a "virtual ego-vehicle" using the data collection vehicle's dimensions. A proposed deviation is deemed valid only if it satisfies two constraints:

- **Dynamic Collision Check:** We perform an Oriented Bounding Box (OBB) intersection test between the virtual ego-vehicle and all labeled dynamic objects at every timestamp.
- **Drivable Region Constraint:** Utilizing the HD Map, we identify free space by rasterizing lane boundaries. We discard trajectories where the vehicle footprint encroaches on non-drivable areas (e.g., sidewalks) or crosses solid lane lines into opposing traffic.

2. Trajectory Parameterization (Gradual Transition). To simulate realistic lane-change maneuvers, we model the trajectory by simultaneously modulating the lateral shift and the vehicle's heading (yaw). Let T be the maneuver duration and S be the total lateral shift. We define a normalized time $\tau = \text{clip}(t/T, 0, 1)$.

- **Lateral Displacement (Half-Cosine Ease):** To ensure smooth acceleration and deceleration, the lateral offset Δy_t follows a half-cosine curve:
$$\Delta y_t = \frac{S}{2} \cdot (1 - \cos(\pi\tau)) \tag{7}$$
This curve ensures zero lateral velocity at both the start and end of the maneuver.

- **Yaw Adaptation (Sine Wave):** Simply translating the camera laterally introduces unnatural sliding artifacts. To mimic realistic steering (turning into the lane and straightening out), we introduce a yaw adjustment $\Delta\psi_t$ modeled by a sine wave:
$$\Delta\psi_t = \psi_{peak} \cdot \sin(\pi\tau) \tag{8}$$
where $\psi_{peak} \approx \min(5^\circ, 2|S|)$. This rotates the view towards the target lane during the shift and returns to the original heading ($\Delta\psi = 0$) upon completion.

- **Pose Synthesis:** The final target pose P'_t is computed by transforming the original pose P_t with the local lateral translation T_{lat} and rotation R_z:
$$P'_t = P_t \cdot T_{lat}(\Delta y_t) \cdot R_z(\Delta\psi_t) \tag{9}$$

D. Self-Supervised Training Data Generation

To enable self-supervised training without paired ground truth, we construct a synthetic training triplet $(\mathcal{S}, \mathcal{C}, V_{gt})$ via two automated pipelines: 3D Asset Curation for geometric scaffolding and Synthetic View Perturbation for trajectory misalignment.

D.1. 3D Asset Curation Pipeline

To construct the library of high-fidelity 3D assets $\mathcal{A}_{obj}$ utilized in our Unified Geometric Scaffold, we implement a rigorous three-stage pipeline consisting of automated filtering, segmentation, and semantic verification.

1. Automated Visibility Filtering. We first parse the raw driving logs to identify potential object candidates. To ensure sufficient multi-view coverage and visibility, we filter objects based on their projection on the camera image plane:

- **Overlap Threshold:** We calculate the intersection between the projected 3D bounding box and the camera canvas. An object is considered visible in a frame only if the intersection ratio exceeds a threshold $\tau_{overlap} = 0.35$.
- **Temporal Stability:** To guarantee robust reconstruction, a candidate must remain visible for at least one-third of the frames within the target video chunk, filtering out transient observations.

2. Instance Segmentation and Pre-processing. For candidates passing the visibility filter, we employ SAM 2 (Ravi et al., 2025) to extract pixel-perfect masks. We utilize the projected 2D bounding boxes as box prompts for the SAM 2 image predictor. The target object is cropped and disconnected small noise regions (area ratio < 0.06) are filtered out. Candidates with a spatial resolution below 120 pixels are discarded to ensure sufficient textural detail.

3. VLM-Based Semantic Verification. Automated segmentation may occasionally yield artifacts, such as truncated vehicles or mis-segmented background elements. We deploy Qwen2.5-VL-7B (Bai et al., 2025) as a semantic verifier to conduct a visual quality inspection.

- **Heuristic Edge Detection:** Before VLM inference, we apply a contour-based heuristic to detect straight edges aligned with image borders. Objects with significant edge truncation are rejected early to save computational costs.
- **Chain-of-Thought Verification:** We design a structured system prompt that requires the VLM to follow a three-step reasoning process: *Observation*, *Analysis*, and *JSON Output*. The model evaluates each asset against four strict criteria:
  1. **Content Check:** Verifying the subject is a vehicle (car, truck, bus) and not background noise.
  2. **Completeness Check:** Ensuring core components (body, roof, wheels) are not truncated or occluded.
  3. **Quality Check:** Rejecting images with severe motion blur or low-light degradation.
  4. **Segmentation Precision:** Checking for jagged edges or the inclusion of detached background elements (e.g., ground patches).

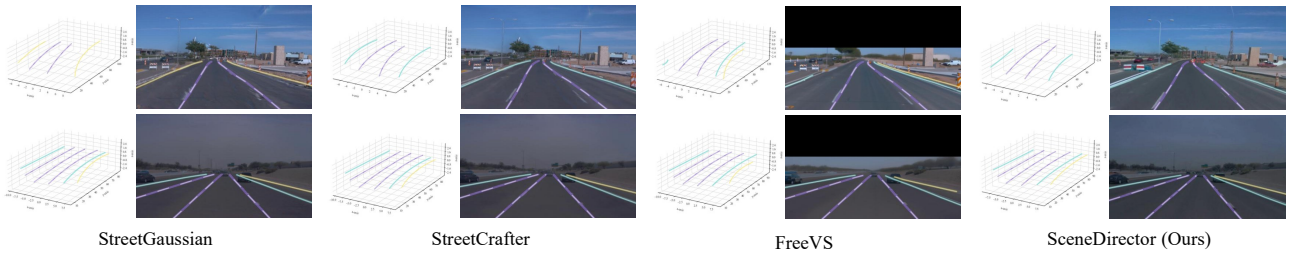

Figure 8. **Qualitative comparison of geometric consistency via lane detection.** We visualize the 3D lane predictions (shown in the 3D grid view on the left of each pair) and their reprojection onto the generated frames (shown on the right).

D.2. Synthetic View Perturbation Function

The perturbation function Φ transforms the ground truth video V_{gt} into a structurally misaligned reference V_{ref} to force the model to rely on the geometric scaffold for alignment. This function Φ simulates trajectory deviations via decoupled temporal and spatial warping strategies.

Temporal Resampling. To simulate longitudinal misalignment (e.g., velocity differences), we apply a speed scaling factor. We resample the video frames via temporal interpolation, effectively compressing or expanding the timeline to mimic acceleration or deceleration relative to the original log.

Spatial Warping. To simulate lateral deviations (e.g., lane changes), we apply view-dependent 2D image warping. The perturbation follows a kinematic trajectory defined by a lateral shift $\Delta x(t)$ and a yaw adjustment $\Delta\psi(t)$.

- **Kinematic Smoothing:** To strictly adhere to physical motion laws, the lateral shift follows a *half-cosine ease curve* to ensure zero velocity at start and end. Simultaneously, the yaw angle follows a *sine wave* (peaking at $\sim 5^\circ$) to mimic the steering banking inherent in lane-change maneuvers.

- **View-Dependent Transforms:** We approximate 3D parallax effects using different 2D warping strategies based on the camera view index:

  1. *Front and Diagonal Views (Affine):* For forward-facing cameras, we apply an Affine transformation combining translation, rotation, and shear. The shear component is crucial to approximate the shift in vanishing points during lateral movement.
  2. *Side Views (Perspective):* For side-facing cameras, affine transforms fail to capture the depth-dependent parallax. We explicitly compute a Perspective transformation (Homography) by warping the image corners, simulating the non-linear distortion of objects passing the field of view.

- **Canvas Filling Constraint:** Warping operations (rotation/shear) typically introduce invalid black borders. To prevent the network from learning trivial shortcuts from these artifacts, we solve for a minimal scaling factor $s_{min} \geq 1.0$ via binary search. This ensures that the warped image coordinates fully cover the target canvas resolution without cropping valid content.

E. Additional Visualization Results

In this section, we provide extensive qualitative results to further demonstrate the superiority of SceneDirector in maintaining structural integrity and photorealism.

E.1. Geometric Consistency via Lane Detection

To explicitly verify whether the synthesized driving scenes preserve valid road topology suitable for downstream perception tasks, we employ a state-of-the-art 3D lane detector, Persformer (Chen et al., 2022), to analyze the generated videos.

Figure 8 presents the visualization of detected 3D lanes and their 2D projections across different methods.

Figure 9. **Qualitative evaluation of editing precision via 3D object detection.** The leftmost column displays the original scene, followed by the editing results from *VACE*, *DriveEditor*, and *SceneDirector*. **Top Row:** A complex combination of object insertion and repositioning (movement). **Bottom Row:** Simultaneous object insertion and replacement. **Legend:** Filled semi-transparent boxes indicate the detection results from the perception model, while hollow wireframe boxes represent the target Ground Truth (GT) layout. **Analysis:** Baseline methods suffer from severe "hallucinations" (numerous false positive boxes in empty areas) and geometric misalignment (detection boxes drifting away from GT). In contrast, *SceneDirector* achieves precise alignment between detections and GT constraints with minimal artifacts, proving its capability to generate functionally valid data for perception training.

E.2. Geometric Precision via Object Detection

Beyond lane topology, we further assess the precision of object manipulation by running a pre-trained 3D object detector on the edited videos. This evaluation serves two purposes: (1) verifying that inserted/modified objects are realistically rendered such that perception algorithms can recognize them, and (2) quantifying the alignment between the user-specified target position (Ground Truth) and the actual generated content. Figure 9 visualizes the detection results across different editing tasks. As observed:

- **Baselines:** *VACE* and *DriveEditor* struggle to strictly adhere to the geometric instructions. We observe significant **geometric drift**, where the detected boxes (filled) deviate noticeably from the target layout (wireframe). Furthermore, they exhibit a high rate of **false positives** (hallucinations), recognizing background artifacts as targets.

- **Ours:** *SceneDirector* demonstrates exceptional geometric fidelity. The detection results align perfectly with the target ground truth boxes, confirming that our method places objects exactly where instructed. Moreover, the clean background—devoid of ghosting artifacts—ensures a low false alarm rate, validating the high quality of our generation.

E.3. Qualitative Results on Fixed Offset Trajectory Editing

Complementing the "Gradual Transition" analysis in the main text (Figure 5), we present comparative results for the **Fixed Offset** setting in Figure 10. This task requires maintaining a constant lateral deviation throughout the sequence, demanding continuous hallucination of occluded regions. As shown in the visualization:

- **Baselines:** Reconstruction-based methods (e.g., *StreetGaussian*, *StreetCrafter*) exhibit characteristic rendering artifacts, such as blurring or streaking in novel views.

- **Ours:** *SceneDirector* overcomes these limitations, generating complete, high-fidelity frames with strictly preserved geometry and consistent lighting, demonstrating robustness across diverse conditions (e.g., night and day).

E.4. Comparative Analysis of Multi-View Trajectory Editing

We extend the analysis from Figure 6 to include baseline comparisons under significant trajectory deviation. As shown in Figure 11, this challenging setting highlights distinct failure modes in existing approaches:

- **Baselines:** Reconstruction-based methods (e.g., StreetGaussian) suffer from severe blurring or voids in disoccluded regions due to limited generative capacity. Pure diffusion methods (e.g., FreeVS), while preserving texture sharpness, fail to maintain structural coherence, leading to noticeable geometric discontinuities (e.g., broken lane lines) across camera boundaries.

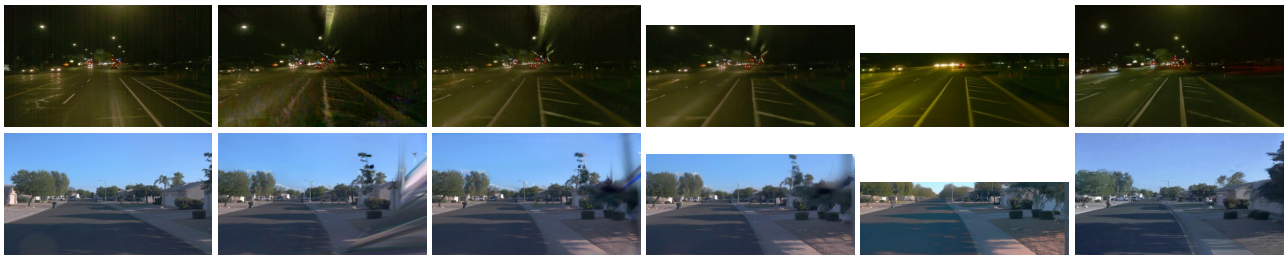

Figure 10. **Qualitative comparison of Fixed Offset trajectory editing.** We compare methods under a constant lateral shift trajectory in both night (top) and day (bottom) scenarios. **Baselines:** *StreetGaussian* and *StreetCrafter* show noticeable rendering distortions. **Ours:** *SceneDirector* successfully reconstructs full-frame, photorealistic video with precise structural alignment, effectively hallucinating plausible details in occluded regions.

Figure 11. **Qualitative comparison of multi-view trajectory editing (Extension of Figure 6).** We compare methods under significant trajectory deviation across adjacent camera views. **Top (Baselines):** Existing methods exhibit characteristic artifacts, including blurring in disoccluded areas (reconstruction-based) and geometric misalignment across frame boundaries (diffusion-based). **Bottom (Ours):** SceneDirector achieves superior performance, maintaining strict geometric consistency across views while synthesizing sharp, realistic textures in newly revealed regions.

- **Ours:** *SceneDirector* effectively bridges these gaps. By leveraging the Unified Geometric Scaffold for strict cross-view alignment and diffusion priors for realistic inpainting, our method generates seamless, photorealistic panoramic views even under large viewpoint shifts.

F. Efficiency Analysis

We assess inference efficiency on NVIDIA A800 GPUs. Note that SceneDirector and VACE-14B utilize 2 GPUs due to memory requirements, while other baselines operate on a single GPU. To ensure a fair comparison, all diffusion-based methods are evaluated using the default sampling steps specified in their respective original implementations.

Object Editing. As shown in Table 7, baselines such as VACE and DriveEditor are restricted to processing a single object within a single view per inference pass. For instance, VACE (830×480) requires $\sim$8 minutes for this limited scope. In contrast, SceneDirector performs unified editing for **4 objects across 5 views** at a higher resolution (1280×704) in a comparable duration ($\sim$15 minutes). Despite the significantly heavier workload and resolution, our parallelized architecture achieves superior throughput by eliminating iterative inference.

Trajectory Editing. Reconstruction-based methods (e.g., StreetGaussian, StreetCrafter) rely on per-scene optimization at high resolutions (1600×1072), leading to variable and lengthy training times. SceneDirector operates as a training-free

feed-forward model, generating novel trajectories directly via inference at 1280×704 without per-scene optimization overhead.

Table 7. **Time cost comparison for generating a 29-frame video clip.** "Scope" indicates the spatial and objective coverage of a single inference pass. Note that *SceneDirector* maintains high efficiency ($\sim$15 min) even while processing multiple objects across surround views at a high resolution (1280×704), whereas baselines like FreeVS operate at significantly lower resolutions (660×380).

TASK	METHOD	# GPUS	RESOLUTION	SCOPE	TIME (MIN)
OBJECT EDITING	VACE	2	830×480	1 VIEW & 1 OBJECT	~ 8
	DRIVEEDITOR	1	1024×576	1 VIEW & 1 OBJECT	~ 5
	SCENEDIRECTOR (OURS)	**2**	$\mathbf{1280 \times 704}$	**5 VIEW & 4 OBJECT**	$\sim$ **15**
TRAJECTORY EDITING	GEM	1	1024×576	1 VIEW	~ 3
	FREEVS	1	660×380	5 VIEW	~ 6
	SCENEDIRECTOR (OURS)	**2**	$\mathbf{1280 \times 704}$	**5 VIEW**	$\sim$ **15**
	STREETCRAFTER	1	1600×1072	1 VIEW	~ 60
	STREETGAUSSIAN	1	1600×1072	5 VIEW	~ 40
	SCENEDIRECTOR+SG (OURS)	**2**	$\mathbf{1600 \times 1072}$	**5 VIEW**	$\sim$ **55**

G. Limitations and Failure Cases

While SceneDirector demonstrates robust performance, we acknowledge certain limitations. First, our reliance on LiDAR-guided depth completion restricts direct applicability in camera-only settings. Second, although our architecture supports variable camera configurations (with embeddings for up to 7 views), the model is currently trained on the 5-camera Waymo setup; generalizing to different sensor layouts (e.g., 6-camera nuScenes) may experience performance drops due to coverage domain gaps, likely requiring fine-tuning.

Regarding failure cases, geometric distortions may appear in thin structures (e.g., traffic sign poles) caused by upstream depth estimation inaccuracies. Additionally, adverse weather conditions, such as rain, can degrade LiDAR data quality, consequently impacting the visual fidelity of the generated video.