OCCLUCAST: CASTING HIDDEN BACKGROUNDS FROM MULTI-VIEW DRIVING VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Removing foreground vehicles from driving videos is a fundamental step toward editable, empty-street reconstruction, yet it remains challenging due to severe occlusions, multi-view consistency requirements, and the need to recover geometrically plausible background content. We present OccluCast, a multi-view temporal diffusion framework for vehicle-free driving-scene inpainting. Given synchronized three-view videos and foreground vehicle masks, OccluCast augments a pretrained video diffusion backbone with a lightweight ViewSync branch that injects same-timestamp multi-view attention and temporal attention to propagate background evidence across cameras and time. An empty-street classifier-free guidance strategy further suppresses foreground-vehicle hallucinations by steering the denoising trajectory away from a jointly learned foreground-reconstruction direction. To lift the inpainted videos into an explicit 3D representation, we perform mesh-guided Gaussian distillation, where a LiDAR-derived background mesh filters the supervision loss so that inpainted content is applied only to genuinely occluded regions while real observations supervise visible surfaces. Experiments on the Waymo benchmark demonstrate that OccluCast produces visually realistic, temporally stable, and geometrically consistent empty-street reconstructions, providing a reliable tool for autonomous driving simulation.

1 INTRODUCTION

Autonomous driving simulators (Dosovitskiy et al., 2017; Yang et al., 2023b; Chen et al., 2025a; Ljungbergh et al., 2024; Jiang et al., 2025a; Zhou et al., 2024a) have attracted increasing attention because they enable safe and cost-effective evaluation of driving algorithms. Conventional simulators (Dosovitskiy et al., 2017; Shah et al., 2017) typically render driving scenes with manually designed CAD assets and rule-based environments. While such systems provide strong controllability, they often exhibit a substantial domain gap from real-world driving data. Recent progress in scene reconstruction offers a promising alternative for high-fidelity autonomous driving simulation by leveraging readily available urban driving images. In particular, advances in neural rendering and 3D Gaussian Splatting (Xie et al., 2023; Yang et al., 2023a; Chen et al., 2026; Yan et al., 2024; Chen et al., 2025b; Ma et al., 2025) have enabled large-scale novel-view synthesis of captured driving scenes with realistic appearance and efficient rendering. Combined with generative priors for improving novel-view synthesis and content completion, existing methods (Yan et al., 2025; Zhao et al., 2025; Yang et al., 2025) can reconstruct recorded urban scenes with increasingly high visual quality. However, most of them remain limited to faithfully replaying the captured environment. Their limited ability to edit scene content poses a major obstacle to evaluating driving algorithms under rare, long-tail, and safety-critical corner cases.

Recovering an empty-street background from recorded driving scenes is a key step toward flexible scene editing. Once foreground vehicles are removed, newly inserted assets can follow more diverse trajectories without being constrained by the original traffic participants. This capability is essential for constructing controllable simulation scenarios for training and evaluating autonomous driving models. Nevertheless, empty-street reconstruction remains challenging. Current Gaussian-based reconstruction methods (Yan et al., 2024; Chen et al., 2025b) struggle to distinguish static vehicles from the persistent scene background. Meanwhile, removing dynamic vehicles exposes regions that were never observed due to occlusion, requiring the model to infer plausible road surfaces, lane

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.

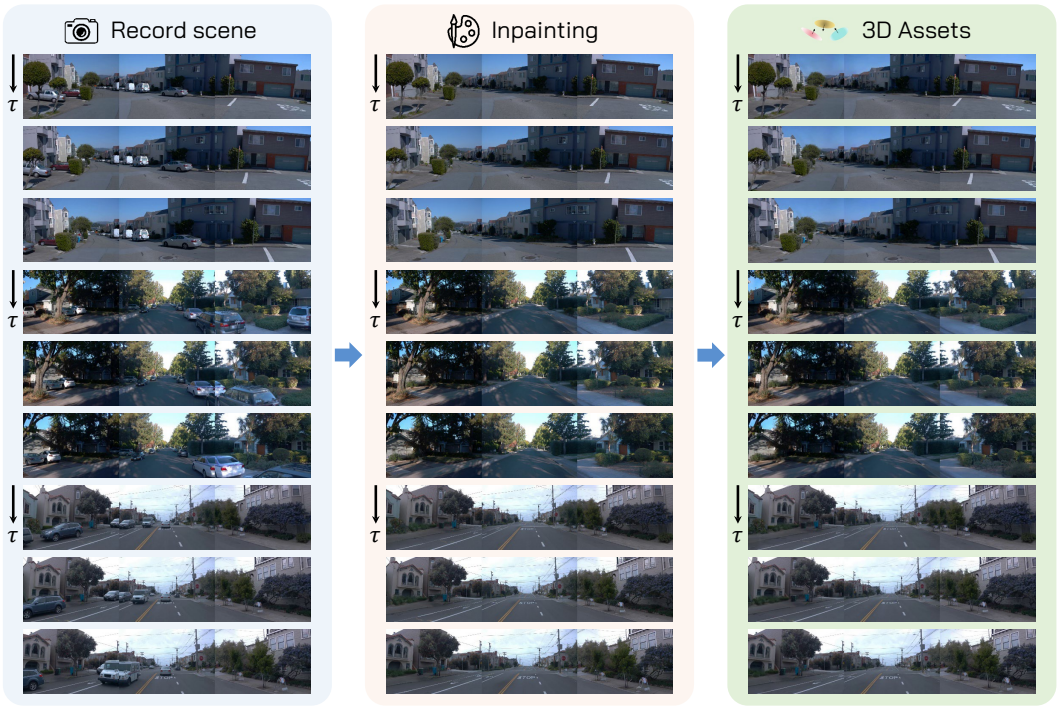


Figure 1: **High-fidelity empty-street reconstruction via OccluCast.** Given recorded driving scenes, OccluCast performs high-quality empty-street inpainting to recover empty-street videos. The completed results are further distilled into editable 3D assets, enabling real-time rendering for autonomous driving simulation.

markings, buildings, and other background structures. These factors make high-fidelity vehicle-free reconstruction substantially more difficult than standard scene reconstruction.

Existing attempts (Xu et al., 2025) typically remove vehicles and feed the resulting holes as masks to image diffusion models on a frame-by-frame basis. Such heuristic masking pipelines are difficult to generalize across diverse driving scenes, object scales, and occlusion patterns. More importantly, independent frame completion often leads to temporal flickering and inconsistent geometry, which are particularly harmful for downstream 3D reconstruction. Recent video inpainting methods (Liang et al., 2025; Wu et al., 2026) improve temporal coherence by modeling consecutive frames, but they still fall short in autonomous driving scenarios. Their generated results are often insufficiently realistic for large-scale street scenes, and most methods ignore the multi-view consistency required by synchronized driving cameras. In addition, the limited number of generated frames makes them difficult to deploy in practical long-range driving sequences, and their compatibility with distillation into explicit 3D representations is rarely considered.

To address these challenges, we present OccluCast, a multi-view temporal diffusion framework for vehicle-free driving-scene inpainting. As qualitatively summarized in Figure 1 and illustrated in Figure 2, OccluCast uses a lightweight ViewSync branch to synthesize empty-street videos that are visually realistic, geometrically coherent across cameras, and stable over long time horizons. Specifically, same-timestamp multi-view attention propagates background evidence across synchronized views, encouraging geometric consistency within each timestamp, while temporal attention enforces long-term coherence across frames and provides reliable observations for subsequent empty-street Gaussian reconstruction. We further introduce an empty-street classifier-free guidance strategy to suppress foreground-vehicle hallucinations during video generation, and mesh-guided Gaussian distillation to leverage LiDAR-derived background geometry and prevent generative artifacts from corrupting the 3D reconstruction. Together, these components provide a scalable and general solution for empty-street reconstruction in autonomous driving simulation, which is crucial for large-scale training and evaluation of driving models.

Our main contributions are summarized as follows: **(i)** We propose a novel video inpainting diffusion framework for scalable empty-street background reconstruction in autonomous driving simu-

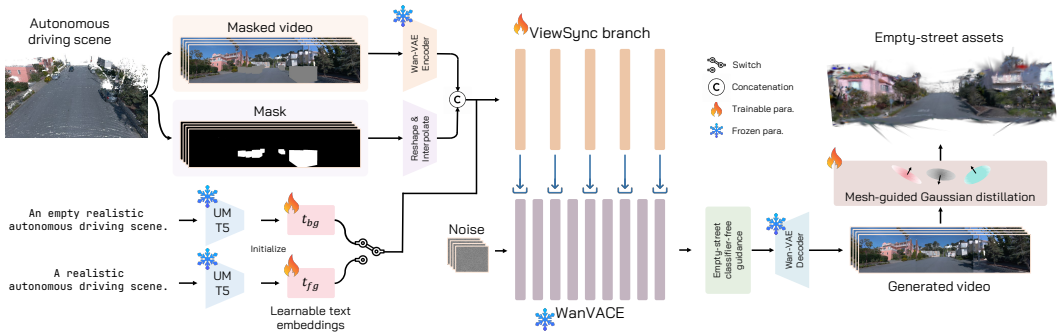


Figure 2: **Pipeline of OccluCast.** Given synchronized multi-view driving videos and foreground masks, OccluCast first synthesizes empty-street videos with the ViewSync branch. By using the empty-street classifier-free guidance strategy, OccluCast suppresses foreground-vehicle hallucinations during generation. The completed videos are then distilled into an explicit 3D Gaussian representation with mesh-guided Gaussian distillation, where LiDAR-derived background geometry prevents conflicts between real observations and generated content.

lation. With the ViewSync branch, our method synthesizes high-quality empty-street videos with both cross-view and temporal consistency. **(ii)** We introduce an empty-street classifier-free guidance strategy and a mesh-guided Gaussian distillation method to jointly reduce generative hallucinations at the video synthesis and 3D reconstruction stages. **(iii)** Extensive experiments on the Waymo benchmark demonstrate that our method outperforms existing state-of-the-art approaches in visual realism and reconstruction quality.

2 Related Work

Neural rendering for autonomous driving simulation. Recent advances in neural rendering have fundamentally changed autonomous driving simulation by enabling high-fidelity, data-driven scene reconstruction directly from sensor logs. Early neural simulators (Yang et al., 2023a; Ost et al., 2021; Turki et al., 2023; Xie et al., 2023; Chen et al., 2025a) mainly relied on neural radiance fields (NeRF) (Mildenhall et al., 2020) to parameterize large-scale urban environments. To overcome the computational bottlenecks and representational limitations of NeRF, recent methods (Chen et al., 2026; Yan et al., 2024; Zhou et al., 2024b; Chen et al., 2025b; Ma et al., 2025; Yang et al., 2024) increasingly adopt 3D Gaussian splatting (3DGS) (Kerbl et al., 2023) for efficient rendering. To alleviate visual artifacts and blurring in unobserved regions under extreme novel views, contemporary frameworks (Wang et al., 2025; Yan et al., 2025; Zhao et al., 2025; Yang et al., 2025) further exploit the strong generative priors of large-scale diffusion models to recover missing details and improve rendering quality. Although these neural rendering and generative frameworks achieve impressive realism when reconstructing recorded data, they are largely constrained to replaying the captured scenes as recorded. This substantially limits the freedom of subsequently inserting safety-critical vehicles and reduces their applicability to evaluating long-tail corner cases in autonomous driving simulators. This limitation motivates us to develop OccluCast, a multi-view temporal diffusion framework that reveals empty-street backgrounds from driving logs and enables flexible downstream scene editing.

Urban scene editing. To test autonomous vehicles in safety-critical scenarios, simulators must support flexible scene editing, making the reconstruction of empty-street backgrounds particularly important. Early works (Wang et al., 2021; Yuan et al., 2020; Philip & Drettakis, 2018; Thonat et al., 2016; Anguelov et al., 2010; Liu et al., 2018; Yu et al., 2018; 2019; Yi et al., 2020; Zhao et al., 2021; Mirzaei et al., 2024) explored street scene editing with single-view or multi-view image inpainting networks. More recently, several works have begun to explore empty-street reconstruction and controllable editing in autonomous driving scenes. DriveEditor (Liang et al., 2025) and RecEdit-Drive (Wu et al., 2026) formulate driving video editing as a unified set of object-level operations, including repositioning, replacement, insertion, and deletion, and use 3D bounding boxes together with reference images to constrain object positions and appearances. However, these methods mainly operate at the 2D video level and do not explicitly reconstruct a freely renderable 3D representation of the empty street. Moreover, how their generated results can be reliably distilled

into a 3D representation remains underexplored. Different from these video editing methods, StreetUnveiler (Xu et al., 2025) explicitly targets empty-street reconstruction. It removes static occluders and generates inpainting regions from rendered alpha maps. Nevertheless, this pipeline relies on semantic segmentation, heuristic mask generation, and external image inpainting models, making it vulnerable to segmentation noise, threshold sensitivity, and generative artifacts. In addition, existing methods are primarily designed for frontal short-trajectory videos, making it difficult to ensure geometric consistency across multiple cameras and long temporal sequences in autonomous driving. In contrast, OccluCast uses the ViewSync branch to synthesize empty-street videos with both long-term stability and cross-view consistency. We further use empty-street classifier-free guidance to suppress foreground-vehicle hallucinations and mesh-guided Gaussian distillation to improve geometric consistency during 3D reconstruction, enabling real-time interactive rendering for closed-loop autonomous driving simulation.

3 METHOD

3.1 PRELIMINARIES

Flow matching. OccluCast builds on Wan2.1 VACE-14B (Wan et al., 2025; Jiang et al., 2025b), which adopts flow matching (Lipman et al., 2023; Esser et al., 2024) for video generation. Given a video latent $\boldsymbol{x}_1$ and Gaussian noise $\boldsymbol{x}_0 \sim \mathcal{N}(0, I)$, rectified flow defines the intermediate latent at timestep $t \in [0, 1]$ as $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$, with target velocity:

$$\boldsymbol{v}_t = \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{x}_1 - \boldsymbol{x}_0. \tag{1}$$

The denoising network is trained to predict this velocity by minimizing:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{c}_{txt}, t} ||\boldsymbol{u}(\boldsymbol{x}_t, \boldsymbol{c}_{txt}, t; \boldsymbol{\theta}) - \boldsymbol{v}_t||^2, \tag{2}$$

where $\boldsymbol{u}(\cdot; \boldsymbol{\theta})$ is the predicted velocity field, $\boldsymbol{c}_{txt}$ denotes the text condition, and $\boldsymbol{\theta}$ denotes trainable parameters.

3.2 DATA CONSTRUCTION

OccluCast is trained with a reconstruction task. Specifically, we occlude foreground objects in driving video sequences and train OccluCast to recover the empty-street background behind the occluders from synchronized three-view frontal images and the corresponding foreground masks. Since existing autonomous driving datasets do not provide ground-truth background scenes with foreground objects removed, we construct a new training set from the widely used Waymo dataset (Sun et al., 2020). For each training sample, we use foreground masks from other scenes and exclude the foreground mask of the current scene, obtaining masks that cover empty regions in the current scene. This strategy trains the model to complete scene backgrounds while matching the mask distribution used at inference time, which improves generation quality during foreground removal.

To obtain accurate foreground masks, we first project the annotated 3D bounding boxes of all foreground objects in the Waymo dataset onto the three frontal cameras. To compensate for missing foreground annotations, we further apply Mask2Former (Cheng et al., 2021) to segment foreground objects in each image and take the union of the segmentation masks and the projected 3D-box masks. This process yields 3,200 RGB video clips and 3,200 mask video clips, which can be freely paired to construct diverse training samples.

3.3 OCCLUCAST ARCHITECTURE

OccluCast augments the pretrained WanVACE with a lightweight ViewSync branch, as shown in Figure 3. The pretrained transformer provides a strong video generation prior, while the newly introduced branch injects foreground-removal conditions and propagates background evidence across synchronized cameras and time. For a three-camera training sample, we denote the RGB videos and masks as

$$\boldsymbol{x} = \{\boldsymbol{x}^L, \boldsymbol{x}^F, \boldsymbol{x}^R\}, \qquad \boldsymbol{m} = \{\boldsymbol{m}^L, \boldsymbol{m}^F, \boldsymbol{m}^R\}, \tag{3}$$

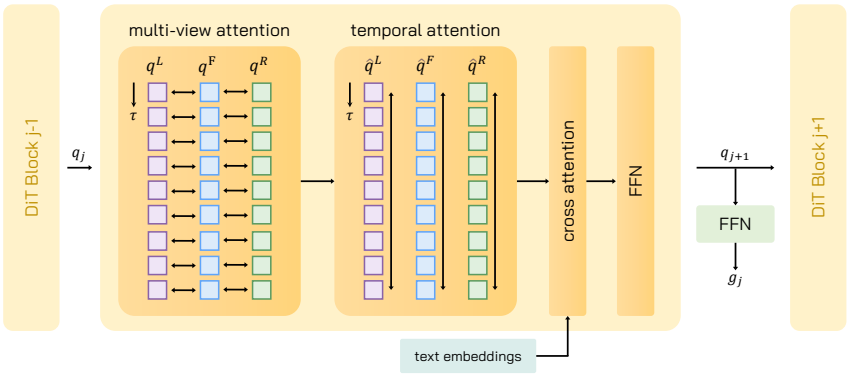


Figure 3: **Structure of ViewSync.** ViewSync is a lightweight branch attached to the frozen WanVACE backbone. It first exchanges information across synchronized cameras through same-timestamp multi-view attention to produce multi-view consistent backgrounds, and then applies temporal attention along each view to generate temporally coherent completions. The refined tokens are finally projected and injected into the backbone blocks.

where L, F, R indicate the left, front, and right cameras. For each camera $c \in \{L, F, R\}$, the VAE encoder $\boldsymbol{\mathcal{E}}$ maps the video $\boldsymbol{x}$ into the latent space $\boldsymbol{z}$. We construct the VACE condition $\boldsymbol{u}$ from the visible context, an empty reactive placeholder, and the downsampled inpainting mask:

$$\boldsymbol{z}^c = \boldsymbol{\mathcal{E}}(\boldsymbol{x}^c), \tag{4}$$

$$\boldsymbol{u}^c = \mathrm{Concat}\big[\boldsymbol{\mathcal{E}}(\boldsymbol{x}^c \odot (1 - \boldsymbol{m}^c)), \boldsymbol{\mathcal{E}}(\boldsymbol{0}), \mathrm{Down}(\boldsymbol{m}^c)\big]. \tag{5}$$

The condition $\boldsymbol{u}^c$ is then tokenized by a 3D patch projection $\mathcal{P}_v$. We further add a learnable camera embedding $\boldsymbol{e}$ to preserve view identity:

$$\boldsymbol{q}_0^c = \mathcal{P}_v(\boldsymbol{u}^c) + \boldsymbol{e}^c, \qquad c \in \{L, F, R\}. \tag{6}$$

The noisy video latent $\boldsymbol{z}$ at diffusion timestep t is processed by the frozen WanVACE backbone. Let $\boldsymbol{h}_0 = \mathcal{P}_x(\boldsymbol{z})$ be the backbone input tokens and $\mathcal{B}_i$ be the i-th pretrained transformer block. We inject the ViewSync branch into every 5th backbone block. For a selected backbone block i, we use j to index its corresponding ViewSync block in injection order, with $j = i/5$ when $i \bmod 5 = 0$. The residual hint $\boldsymbol{g}_j$ is then added to the backbone feature:

$$\boldsymbol{h}_{i+1} = \mathcal{B}_i(\boldsymbol{h}_i) + \alpha_j \boldsymbol{g}_j \mathbb{I}[i \bmod 5 = 0]. \tag{7}$$

Here, α_j controls the strength of the j-th residual injection. This residual design keeps the pretrained generation pathway intact and confines task-specific adaptation to the ViewSync branch.

Each ViewSync block indexed by j is designed to capture two complementary forms of consistency. First, it performs same-timestamp multi-view attention. For each frame index τ, tokens from the three cameras are concatenated and attended under a view mask $\boldsymbol{M}_{\mathrm{view}}$:

$$\boldsymbol{s}_\tau = \mathrm{Concat}_{c \in \{L,F,R\}}\big[\boldsymbol{q}^c_{j,\tau,1:P}\big], \tag{8}$$

$$\hat{\boldsymbol{s}}_\tau = \boldsymbol{s}_\tau + \mathrm{Attn}_{\mathrm{view}}\big(\mathrm{AdaLN}(\boldsymbol{s}_\tau; t), \boldsymbol{M}_{\mathrm{view}}\big). \tag{9}$$

where P denotes the number of spatial patches per frame and $\boldsymbol{M}_{\mathrm{view}} \in \{0,1\}^{3P \times 3P}$ is a block-structured view mask over the three camera segments. The mask retains the intra-camera blocks and the two adjacent cross-camera pairs, left–front and front–right, while blocking the direct left–right interaction. This follows the camera layout, in which the two side cameras share almost no common field of view. The attended sequence is then split back into camera-specific tokens:

$$\{\hat{\boldsymbol{q}}^L_{j,\tau,1:P}, \hat{\boldsymbol{q}}^F_{j,\tau,1:P}, \hat{\boldsymbol{q}}^R_{j,\tau,1:P}\} = \mathrm{Split}(\hat{\boldsymbol{s}}_\tau). \tag{10}$$

This module allows overlapping camera views to exchange background cues, which encourages cross-view geometric consistency. Second, after reshaping the updated tokens back to each camera stream, the block applies temporal attention to enforce coherent completion across frames:

$$\tilde{\boldsymbol{q}}^c_j = \hat{\boldsymbol{q}}^c_j + \mathrm{Attn}_{\mathrm{time}}\big(\mathrm{AdaLN}(\hat{\boldsymbol{q}}^c_j; t)\big). \tag{11}$$

The updated tokens are then conditioned on text features $\boldsymbol{y}$ and refined by a feed-forward network. Finally, the refined control tokens are projected to the backbone hidden space to produce $\boldsymbol{g}_j$ for residual injection:

$$\bar{\boldsymbol{q}}_j = \tilde{\boldsymbol{q}}_j + \mathrm{CrossAttn}\big(\mathrm{LN}(\tilde{\boldsymbol{q}}_j), \boldsymbol{y}\big), \tag{12}$$

$$\boldsymbol{q}_{j+1} = \bar{\boldsymbol{q}}_j + \mathrm{FFN}\big(\mathrm{AdaLN}(\bar{\boldsymbol{q}}_j; t)\big), \tag{13}$$

$$\boldsymbol{g}_j = W_{\mathrm{out}} \boldsymbol{q}_{j+1}. \tag{14}$$

This projection-and-injection design separates conditional reasoning from video generation. The ViewSync branch can focus on aggregating mask, view, and temporal cues, while the frozen backbone remains responsible for synthesizing realistic video content. Injecting $\boldsymbol{g}_j$ as a residual hint also makes the conditioning non-destructive: when the control signal is uncertain, the backbone prior can still preserve plausible texture and motion, which is important for completing severely occluded street regions. During training, we freeze the pretrained WanVACE backbone and optimize only the added ViewSync blocks and camera embeddings. This parameter-efficient adaptation preserves the visual prior of the foundation model while specializing its control pathway for multi-view empty-street synthesis.

3.4 Empty-street classifier-free guidance

Although fine-tuning adapts the pretrained video model to empty-street synthesis, the original generative prior still tends to hallucinate vehicles inside large masked regions. This behavior is expected because realistic driving videos frequently contain foreground traffic participants, and the model may interpret a large missing region as a plausible location for a vehicle. To suppress such foreground artifacts, we introduce an empty-street classifier-free guidance strategy (Ho & Salimans, 2022). Instead of learning only the target empty-street distribution, we jointly learn two conditional directions: one for recovering empty-street backgrounds and the other for reconstructing foreground vehicles. At inference time, we move the denoising direction away from the foreground-vehicle direction.

We define two learnable text embeddings, $\boldsymbol{t}_{\mathrm{bg}}$ and $\boldsymbol{t}_{\mathrm{fg}}$, where $\boldsymbol{t}_{\mathrm{bg}}$ represents the empty-street background direction and $\boldsymbol{t}_{\mathrm{fg}}$ represents the foreground-vehicle direction. They are initialized by UMT5 (Chung et al., 2023) using the prompts "An empty realistic autonomous driving scene." and "A realistic autonomous driving scene.", respectively. During training, the empty-street branch uses masks sampled from other scenes to occlude empty regions in the current scene, and is conditioned on $\boldsymbol{t}_{\mathrm{bg}}$. The foreground branch uses the current-scene foreground mask to occlude vehicles, and is conditioned on $\boldsymbol{t}_{\mathrm{fg}}$. Thus, the model observes both how to complete empty-street backgrounds and how the same backbone would move toward generating foreground vehicles:

$$\hat{\boldsymbol{v}}_{\mathrm{bg}} = \boldsymbol{u}(\boldsymbol{z}_t, \boldsymbol{u}^c_{\mathrm{bg}}, \boldsymbol{t}_{\mathrm{bg}}, t; \boldsymbol{\theta}), \tag{15}$$

$$\hat{\boldsymbol{v}}_{\mathrm{fg}} = \boldsymbol{u}(\boldsymbol{z}_t, \boldsymbol{u}^c_{\mathrm{fg}}, \boldsymbol{t}_{\mathrm{fg}}, t; \boldsymbol{\theta}). \tag{16}$$

Here, $\boldsymbol{u}^c_{\mathrm{bg}}$ and $\boldsymbol{u}^c_{\mathrm{fg}}$ denote the VACE conditions built from the empty-region mask and the foreground-vehicle mask.

During inference, for each denoising timestep, we perform two forward passes with the empty-street and foreground conditions. The final guided velocity is obtained by subtracting the foreground direction from the empty-street direction:

$$\hat{\boldsymbol{v}}_{\mathrm{cfg}} = \hat{\boldsymbol{v}}_{\mathrm{bg}} + \lambda_{\mathrm{cfg}}(\hat{\boldsymbol{v}}_{\mathrm{bg}} - \hat{\boldsymbol{v}}_{\mathrm{fg}}), \tag{17}$$

where λ_{cfg} controls the strength of foreground suppression. This guidance explicitly penalizes denoising updates that resemble foreground-vehicle reconstruction, while preserving the texture and illumination priors learned from the pretrained video model. As a result, the model is encouraged to fill large occluded regions with road surfaces, lane markings, and static structures rather than vehicle-like artifacts.

3.5 Mesh-guided Gaussian distillation

To obtain a 3D empty-street representation, we distill the generated empty-street videos into 3D Gaussians (Kerbl et al., 2023) while retaining supervision from the original observations. Specifically, we optimize two coupled renderings during reconstruction. Rendering all Gaussians $\mathcal{G}$ is

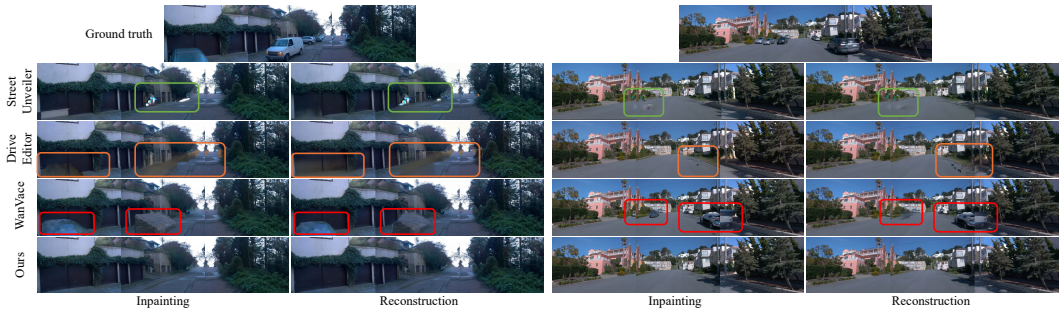


Figure 4: **Qualitative comparison on the Waymo (Sun et al., 2020) dataset.** Compared with existing baselines, OccluCast removes foreground vehicles more completely and generates cleaner empty-street backgrounds with fewer hallucinated artifacts. The distilled 3D representation further preserves sharper geometry and more consistent appearance for novel-view rendering.

supervised by the original video $\boldsymbol{x}$, which preserves the recorded scene content and stabilizes geometry in observed regions. Rendering only the background Gaussians $\mathcal{G}_{\text{bg}}$ is supervised by the inpainted empty-street video $\hat{\boldsymbol{x}}_{\text{bg}}$, which transfers the completed vehicle-free appearance into the explicit 3D representation:

$$\hat{\boldsymbol{x}} = \mathcal{R}(\mathcal{G}), \qquad \hat{\boldsymbol{x}}^{3D}_{\text{bg}} = \mathcal{R}(\mathcal{G}_{\text{bg}}), \tag{18}$$

$$\mathcal{L}_{\text{rgb}} = ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_1 + ||\boldsymbol{m}_{\text{sup}} \odot (\hat{\boldsymbol{x}}^{3D}_{\text{bg}} - \hat{\boldsymbol{x}}_{\text{bg}})||_1. \tag{19}$$

Here, $\mathcal{R}(\cdot)$ denotes differentiable Gaussian rendering, and $\boldsymbol{m}_{\text{sup}}$ denotes the valid supervision mask for the inpainted empty-street frames.

Directly supervising $\mathcal{G}_{\text{bg}}$ with all inpainted pixels can be harmful because some masked pixels correspond to background surfaces that are actually observed in other frames or cameras. In these regions, the generated content may conflict with real multi-view observations, causing blurred textures and inconsistent geometry after optimization. To avoid this conflict, we construct a geometry-aware mask from LiDAR observations. For each frame, we project the LiDAR point cloud onto the camera plane, retain camera-visible points, aggregate them across frames, and remove foreground LiDAR points using foreground object masks and annotations. We then apply VDBFusion (Vizzo et al., 2022) to the aggregated background point cloud and remove floaters, producing a background mesh $\mathcal{M}_{\text{bg}}$ that covers regions with reliable visible-background evidence.

During empty-street supervision, we project $\mathcal{M}_{\text{bg}}$ back to each camera and exclude the projected mesh region from the inpainting supervision mask:

$$\boldsymbol{m}_{\text{sup}} = \boldsymbol{m}_{\text{fg}} \odot \big(1 - \Pi(\mathcal{M}_{\text{bg}})\big), \tag{20}$$

where $\Pi(\cdot)$ denotes projection to camera, and $\boldsymbol{m}_{\text{fg}}$ is the foreground removal mask. This mesh-guided filtering lets real observations supervise visible background surfaces, while using inpainted frames only for genuinely occluded regions. As a result, the reconstruction avoids conflicts between ground-truth observations and generated content, leading to sharper background Gaussians and more coherent empty-street geometry.

4 Experiments

4.1 Experimental setup

Baselines. We compare OccluCast with three representative baselines. StreetUnveiler (Xu et al., 2025) targets empty-street reconstruction with object removal and inpainting-based pseudo labels. DriveEditor (Liang et al., 2025) is evaluated under its deletion setting as a video-level foreground removal baseline. We also include WanVACE (Jiang et al., 2025b), the pretrained inpainting backbone of our method, without ViewSync, empty-street classifier-free guidance, or mesh-guided Gaussian distillation. This comparison separates the gains from the pretrained generative prior and the task-specific designs introduced in OccluCast.

Evaluation metrics. We evaluate both generated empty-street videos and rendered reconstructions against original videos. We report FID (Heusel et al., 2017) for frame-level visual realism and

Table 1: **Comparison on Waymo (Sun et al., 2020).** We evaluate both generated empty-street videos and rendered reconstruction results against the original videos. MV Cons. denotes the normalized mask-aware LPIPS overlap consistency score. Best and second-best results among compared methods are shown in **bold** and underlined, respectively.

Methods	Generation			Reconstruction		
	FID↓	FVD↓	MV Cons.↑	FID↓	FVD↓	MV Cons.↑
StreetUnveiler (Xu et al., 2025)	119.77	621.21	69.57	120.11	433.30	81.34
DriveEditor (Liang et al., 2025)	110.77	508.90	54.23	116.44	487.23	75.38
WanVACE (Jiang et al., 2025b)	110.93	494.36	63.50	113.15	492.96	72.46
OccluCast (Ours)	**79.06**	**261.87**	**79.42**	**82.92**	**231.86**	**88.26**

Table 2: **Ablation study on Waymo (Sun et al., 2020).** We evaluate the contributions of ViewSync, empty-street classifier-free guidance, and mesh-guided Gaussian distillation.

Settings	Generation			Reconstruction		
	FID↓	FVD↓	MV Cons.↑	FID↓	FVD↓	MV Cons.↑
w/o ViewSync	88.31	312.95	71.42	93.45	307.79	79.63
w/o empty-street CFG	101.22	340.95	70.29	108.93	364.67	76.57
w/o mesh-guided distillation	**79.06**	**261.87**	**79.42**	83.12	235.59	86.95
OccluCast (Ours)	**79.06**	**261.87**	**79.42**	**82.92**	**231.86**	**88.26**

FVD (Unterthiner et al., 2019) for video-level temporal consistency. To measure multi-view consistency, we introduce a mask-aware LPIPS (Zhang et al., 2018)-based overlap score for adjacent camera views. For each scene, we sweep candidate overlap widths on the ground-truth adjacent views and restrict each comparison to the bounding box of the intersection between the overlap region and the union of the inpainting masks. For each candidate width, we compute LPIPS on matched ground-truth adjacent-view pairs and on randomly shuffled ground-truth pairs, where the latter serve as an inconsistent-view reference. We select the width that maximizes the separation between shuffled-pair and matched-pair LPIPS. The selected width and random reference are then fixed for all methods. Finally, each generated adjacent-view pair is normalized into a consistency percentage, where 100% corresponds to ground-truth adjacent-view consistency and 0% corresponds to the random-pair baseline.

Implementation details. OccluCast is implemented on Wan2.1 VACE-14B (Jiang et al., 2025b). We freeze the Wan backbone and optimize only the 3B-parameter ViewSync blocks and camera embeddings, which are injected into every 5th Wan-DiT block. We train on 49-frame clips at 480×720 resolution using Adam (Kingma & Ba, 2017) with learning rate 1×10^{-5}, batch size 8, and 1000 warmup iterations. Training runs for 20,000 iterations on 8 NVIDIA H200 GPUs. At inference, we use 30 denoising steps with $\lambda_{\text{cfg}} = 2.5$. For 3D extraction, we follow Street Gaussians (Yan et al., 2024) to optimize the Gaussian scene representation. We disable LiDAR depth supervision in completed regions because generated vehicle-free content lacks reliable depth ground truth. Training a standard 49-frame sequence takes about 1.5 hours on a single NVIDIA RTX 4090 GPU.

4.2 Comparisons with the state-of-the-art

As shown in Table 1, OccluCast consistently outperforms existing baselines on both video generation and 3D reconstruction. The qualitative comparisons in Figure 4 further show that our method produces cleaner empty-street completions with fewer foreground remnants and more coherent background structures. These results indicate that the proposed framework improves not only perceptual realism, measured by FID, but also video-level stability and cross-view consistency, measured by FVD and MV Cons., respectively.

For empty-street video generation, OccluCast achieves the best FID and FVD while maintaining the strongest multi-view consistency. DriveEditor and WanVACE rely mainly on video-level inpainting and therefore lack explicit communication across synchronized cameras. As a result, their completions can be locally plausible within a single view but inconsistent across adjacent views or unstable over time. In contrast, the ViewSync branch performs same-timestamp multi-view attention and temporal attention, allowing background evidence to be shared across cameras and propagated

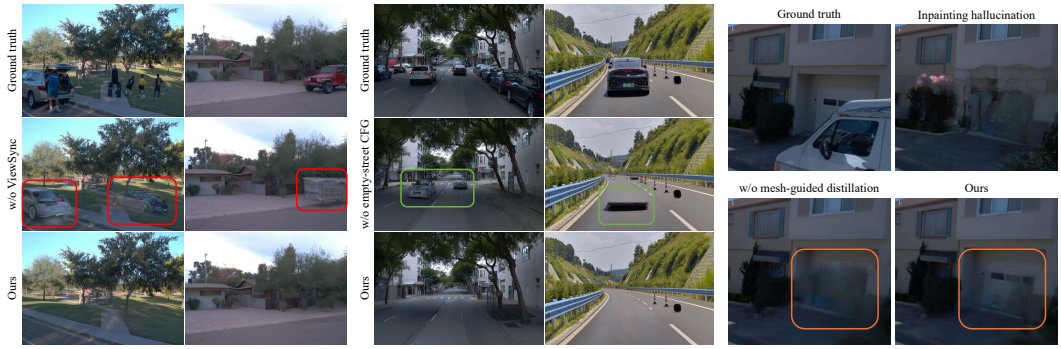


Figure 5: **Ablation study on architectural designs.** The visual results show the effects of removing ViewSync, empty-street classifier-free guidance, and mesh-guided Gaussian distillation. ViewSync improves cross-view and temporal consistency, empty-street classifier-free guidance suppresses foreground-vehicle hallucinations, and mesh-guided Gaussian distillation reduces reconstruction artifacts.

along the sequence. This explains the improvements in visual quality, temporal coherence, and MV Cons. StreetUnveiler obtains competitive multi-view consistency because its masks are derived from a geometry-aware reconstruction process, but its heuristic mask generation and image-level completion limit frame-level fidelity and temporal stability.

The advantage becomes more evident after distilling empty-street videos into a 3D representation. OccluCast achieves the best reconstruction scores across all metrics, indicating that generated videos provide cleaner and more stable supervision for Gaussian reconstruction. Compared with methods that rely on inpainted pseudo labels in masked regions, mesh-guided Gaussian distillation jointly uses original observations and completed empty-street views while suppressing conflicts between visible ground truth and generated content. This design reduces blurry geometry and appearance inconsistency during optimization, producing a sharper and more coherent empty-street representation for downstream autonomous driving simulation.

4.3 Ablation Study

Table 2 and Figure 5 ablate the main task-specific designs in OccluCast. Without ViewSync, the WanVACE backbone is fine-tuned without the task-specific modules, degrading visual quality, temporal stability, and multi-view consistency. This gap shows that ViewSync is crucial for synchronizing adjacent views and preserving temporal coherence in multi-camera driving videos. Empty-street classifier-free guidance further improves the generation direction by suppressing the foreground-vehicle prior inherited from driving videos. Removing this guidance degrades both generation and reconstruction, indicating that cleaner empty-street completions provide more reliable supervision for Gaussian optimization. Mesh-guided Gaussian distillation mainly affects the reconstruction stage: the generation metrics remain unchanged, while reconstruction quality and multi-view consistency drop without it. This verifies that the mesh prior helps avoid conflicts between visible ground truth and generated completions during 3D reconstruction.

5 Conclusion

We presented OccluCast, a multi-view temporal diffusion framework for recovering empty-street backgrounds from synchronized driving videos. OccluCast augments a pretrained video diffusion backbone with a lightweight ViewSync branch that combines same-timestamp multi-view attention and temporal attention, producing empty-street videos that are visually realistic, temporally stable, and geometrically coherent. An empty-street classifier-free guidance strategy further suppresses foreground-vehicle hallucinations by steering denoising away from a jointly learned foreground-reconstruction direction. To lift the generated videos into an explicit 3D representation, we use mesh-guided Gaussian distillation, where a LiDAR-derived background mesh ensures that inpainted content is used only for genuinely occluded regions. Experiments on the Waymo benchmark demonstrate that OccluCast outperforms existing methods in both visual realism and reconstruction quality, providing a robust tool for empty-street reconstruction in autonomous driving simulation.

AI USE STATEMENT

In this work, we used generative AI tools for translating the authors' drafts of this manuscript into English, and polishing the resulting English text. We have not used generative AI tools for any other task that requires mandatory disclosure. All research ideas and hypotheses, the problem formulation, the method design, the theoretical formulations, the experimental design, the data collection and processing, the implementation of the proposed method, and the analysis and interpretation of results were carried out entirely by the authors. No synthetic data were generated with AI tools, and no AI tool was used to produce experimental results for this work. We also did not use AI tools to write the code used in this work, nor to generate or edit any figures, tables, or references. We have reviewed all AI-assisted text, checked the translated text against our original drafts for fidelity, and verified that it accurately represents our scientific contributions and claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our work. To this end, we will release the full implementation and relevant scripts upon the final acceptance of the paper. This timeline allows us to carefully clean and document the code for ease of use. The construction of the training data from the publicly available Waymo Open Dataset, including the generation of the foreground masks, is described in Section 3.2. The architecture of OccluCast and the ViewSync branch are detailed in Section 3.3, the empty-street classifier-free guidance in Section 3.4, and the mesh-guided Gaussian distillation in Section 3.5. All experimental settings are specified in Section 4, including the compared baselines, the definition of the evaluation metrics, and the training and inference hyperparameters; the ablation configurations are summarized in Table 2. Our method builds only on publicly available datasets and pretrained models, so all results can be verified prior to code release.

REFERENCES

Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stephane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 2010.

Yurui Chen, Junge Zhang, Ziyang Xie, Wenye Li, Feihu Zhang, Jiachen Lu, and Li Zhang. S-nerf++: Autonomous driving simulation via neural reconstruction and generation. *PAMI*, 2025a.

Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *IJCV*, 2026.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. In *ICLR*, 2025b.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining, 2023.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Junzhe Jiang, Nan Song, Jingyu Li, Xiatian Zhu, and Li Zhang. Realengine: Simulating autonomous driving in realistic context, 2025a.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing, 2025b.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *TOG*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Yiyuan Liang, Zhiying Yan, Liqun Chen, Jiahuan Zhou, Luxin Yan, Sheng Zhong, and Xu Zou. Driveeditor: A unified 3d information-guided framework for controllable object editing in driving scenes. In *AAAI*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Guilin Liu, Kevin J Shih, Ting-Chun Wang, Fitsum A Reda, Karan Sapra, Zhiding Yu, Andrew Tao, and Bryan Catanzaro. Partial convolution based padding, 2018.

William Ljungbergh, Adam Tonderski, Joakim Johnander, Holger Caesar, Kalle Åström, Michael Felsberg, and Christoffer Petersson. Neuroncap: Photorealistic closed-loop safety testing for autonomous driving. In *ECCV*, 2024.

Zipei Ma, Junzhe Jiang, Yurui Chen, and Li Zhang. Béziergs: Dynamic urban scene reconstruction with bézier curve gaussian splatting. In *ICCV*, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Ashkan Mirzaei, Riccardo De Lutio, Seung Wook Kim, David Acuna, Jonathan Kelly, Sanja Fidler, Igor Gilitschenski, and Zan Gojcic. Reffusion: Reference adapted diffusion models for 3d scene inpainting, 2024.

Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021.

Julien Philip and George Drettakis. Plane-based multi-view inpainting for image-based rendering in large scenes. In *I3D*, 2018.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *FSR*, 2017.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

Theo Thonat, Eli Shechtman, Sylvain Paris, and George Drettakis. Multi-view inpainting for image-based scene editing and rendering. In *3DV*, 2016.

Haithem Turki, Jason Y. Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes, 2023.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019.

Ignacio Vizzo, Tiziano Guadagnino, Jens Behley, and Cyrill Stachniss. Vdbfusion: Flexible and efficient tsdf integration of range sensor data. *Sensors*, 2022.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025.

Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. In *ICLR*, 2025.

Yifan Wang, Andrew Liu, Richard Tucker, Jiajun Wu, Brian L. Curless, Steven M. Seitz, and Noah Snavely. Repopulating street scenes. In *CVPR*, 2021.

Yipeng Wu, Xin Wang, Chenghan Yang, Chong Wang, Dongdong Wu, Wanchao Su, Hengshuang Zhao, Wei Feng, Kairui Yang, and Di Lin. Recedit-drive: 3d reconstruction-guided spatiotemporal video editing for autonomous driving scenes. In *CVPR*, 2026.

Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. In *ICLR*, 2023.

Jingwei Xu, Yikai Wang, Yiqun Zhao, Yanwei Fu, and Shenghua Gao. 3d streetunveiler with semantic-aware 2dgs - a simple baseline. In *ICLR*, 2025.

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 *ECCV*, 2024.

Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, and Sida Peng. Streetcrafter: Street view synthesis with controllable video diffusion models. In *CVPR*, 2025.

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

Jiawei Yang, Jiahui Huang, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Apoorva Sharma, Maximilian Igl, Peter Karkus, Danfei Xu, Boris Ivanovic, Yue Wang, and Marco Pavone. Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes, 2024.

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

Zeyu Yang, Zijie Pan, Yuankun Yang, Xiatian Zhu, and Li Zhang. Driving scene synthesis on free-form trajectories with generative prior. In *ICCV*, 2025.

Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *CVPR*, 2020.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019.

Zefeng Yuan, Hengyu Li, Jingyi Liu, and Jun Luo. Multiview scene image inpainting based on conditional generative adversarial networks. *TIV*, 2020.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In *CVPR*, 2025.

Zibo Zhao, Wen Liu, Yanyu Xu, Xianing Chen, Weixin Luo, Lei Jin, Bohui Zhu, Tong Liu, Binqiang Zhao, and Shenghua Gao. Prior based human completion. In *CVPR*, 2021.

Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving, 2024a.

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