# FrozenDrive: Zero-Shot Text-Guided Driving Scene Generation and Data Augmentation with Parameter-Free Frozen Diffusion Model

Yuhwan Jeong*, Hyeonseong Kim*, Daehyun We*, Seonkyu Song*, Jinnyeong Yang*, Hyun-Kurl Jang, Youngho Yoon, and Kuk-Jin Yoon

KAIST, Visual Intelligence Lab.
{jeongyh98, brian617, greathyun, sok9854, jinnyeong6118, jhg0001, dudgh1732, kjyoon}@kaist.ac.kr

**Abstract.** Synthetic data for autonomous driving is surging, powered by diffusion models that promise scalable scene generation. Yet key obstacles remain, since multi-view and temporal consistency often requires backbone fine-tuning or added layers, which can erode pretrained knowledge and weaken text alignment. Models also stay close to the training distribution, struggling under adverse weather and unseen configurations, while fidelity favors frequent over rare classes. We address these gaps with FrozenDrive, a controllable generative framework that preserves pretrained diffusion knowledge while achieving strong consistency. FrozenDrive conditions on driving-stack signals and text prompts, and expands the context of frozen self-attention across views and frames to promote cross-view alignment and temporal coherence in one pass, without trainable spatio-temporal modules in the diffusion backbone. An object-focused constraint further improves fidelity for rare categories. Without weather- or scene-specific fine-tuning, FrozenDrive synthesizes globally coherent multi-view driving scenes from text and surpasses prior baselines under adverse and rare conditions. On nuScenes, FrozenDrive-augmented data improves AD model performance, especially at night and in rain, demonstrating strong robustness with scenario-targeted data.

**Keywords:** Multi-view generation · Autonomous driving · Data augmentation
Project page: https://frozendriveeccv.github.io/

## 1 Introduction

Training autonomous driving models [10, 11, 18, 27, 32, 71, 83, 88] is constrained by the cost of collecting large-scale datasets. This challenge becomes even more severe when accurate labels require multi-sensor calibration and fine-grained delineation of complex scenes. Moreover, acquiring data under adverse weather or in rare situations is intrinsically difficult: collection is error-prone, annotation

* Equal contribution.

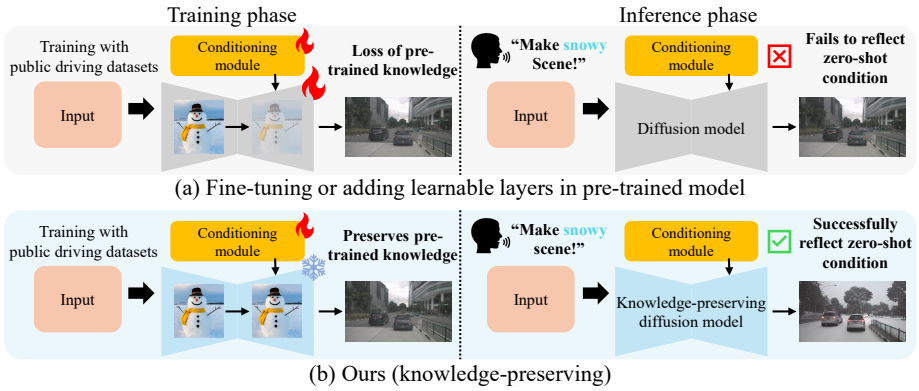



**Fig. 1:** (a) Fine-tuning or adding learnable layers lose pre-trained priors and fails to reflect unseen text prompts while training the model. (b) By completely freezing the pretrained model without introducing any additional parameters, our method preserves its rich generative priors, enabling successful zero-shot text-guided generation.

is even more labor-intensive, and genuinely rare events offer few opportunities for capture, making repeated observations of the same scenario inherently difficult. As a result, real-world corpora are biased toward frequent, benign conditions, and models trained on them exhibit brittle behavior in the long tail.

Recent advances in generative modeling suggest that synthetic data can amortize annotation effort by producing diverse samples from a single label [29, 39, 91]. Mainstream methods for autonomous driving scene generation achieve high visual quality under everyday conditions [14–16, 19, 79, 84, 93, 105, 115]. It synthesizes photorealistic driving scenes by conditioning on 2D spatial controls [104, 106] derived from driving-stack signals—BEV layouts, object bounding boxes, and occupancy maps [89]. To encourage natural object motion, many methods reference features from the previous frame or process the entire sequence. Regardless of the conditioning used, these methods ultimately build on the powerful generative capacity of pretrained diffusion models such as SD [63], SVD [4], and DiT [111], which are adapted to autonomous driving datasets. As a result, prior works have enabled highly controllable and realistic scene generation. Furthermore, by explicitly refining text–scene alignment during training, these methods afford fine-grained, semantically consistent editing of the scene's weather and environment to seen conditions, *e.g.*, transforming an originally clear setting into rain or night conditions observed in the training dataset.

These works, in order to support multi-camera rigs and temporal dynamics, either (i) generate each view largely independently with weak feature sharing [73, 115], or (ii) introduce cross-view and temporal attention via additional layers fine-tuned on top of a pretrained backbone [16, 84, 89, 93].

While these designs improve consistency under seen conditions, they remain bounded by the training distribution. Models trained on major benchmarks [6,70] perform well in standard daytime scenarios but degrade in underrepresented set-

tings. This degradation mainly stems from extensive backbone fine-tuning on datasets with significant coverage gaps. Consequently, the rich priors of the original pretrained model for adverse or unseen scenarios (*e.g.*, heavy snowfall) are largely overwritten or catastrophically forgotten. Furthermore, such reliance on full backbone updates not only compromises the pretrained text-to-image alignment but also fails to preserve per-object fidelity, particularly for rare classes.

To address these limitations, we propose FrozenDrive, a generative framework that preserves pretrained diffusion knowledge while enabling precise control over multi-view imagery under adverse or unseen scenarios. To achieve consistency without trainable spatio-temporal modules in the backbone, FrozenDrive restructures frozen self-attention context through multi-view inflated self-attention and temporal reference self-attention. It concatenates view features to promote circular cross-view alignment and reuses preceding-frame features as references to encourage temporal coherence in a single pass. **Yet the most consequential difference lies elsewhere: we neither introduce additional parameters nor update the pretrained weights.** By keeping the diffusion backbone frozen and parameter-free, FrozenDrive preserves pretrained text–image alignment and avoids knowledge erosion, retaining zero-shot text guidance. In contrast, prior approaches [16,84,89,93] typically strengthen consistency by tuning the backbone or adding trainable modules, which can weaken alignment and cause local or superficial edits. As shown in Fig. 1, such methods [93] often struggle to compose plausible snow scenes unseen during training, whereas FrozenDrive produces coherent results without weather-specific fine-tuning.

To further secure per-object fidelity, we impose an object-focused constraint for rare objects. This shifts learning toward underrepresented categories, improving their generation quality rather than privileging only frequent classes. Together, our design secures robust text alignment and multi-camera consistency, yielding diverse, high-quality street views. With these components, our synthetic data captures fine-grained nuScenes details and achieves state-of-the-art perception and planning performance evaluated with UniAD [27].

Moreover, we leverage text prompts to specify a wide range of driving scenarios, from common conditions to rare, hard-to-collect cases. Consequently, **FrozenDrive enables scenario-targeted data generation that strengthens downstream robustness**: AD models trained with FrozenDrive-augmented data achieve superior perception and planning performance under adverse conditions (*e.g.*, night and rain) on nuScenes, outperforming prior methods.

Our main contributions can be summarized as follows:

- We introduce FrozenDrive, a knowledge-preserving driving scene generator that promotes multi-view and temporal consistency by restructuring frozen self-attention context in a parameter-free pretrained diffusion backbone, enhanced with object-balanced constraints and structured conditioning signals.
- We enable text-guided compositional control over diverse scene factors, time of day, weather conditions, and rare environments, supporting zero-shot synthesis of unseen combinations.

– We demonstrate improved synthesis quality and superior transfer to downstream autonomous driving tasks under both common and adverse weather.

## 2 Related works

### 2.1 Diffusion models

Diffusion models [24, 25, 68, 69] synthesize images via iterative denoising but are memory-prohibitive at high resolutions in pixel space. Latent diffusion models [63] address this by operating in a latent space, enabling efficient, high-fidelity generation, and large-scale text-conditioned pretraining further broadens their applications. Recent extensions to video explicitly model temporal consistency for coherent synthesis [4, 22, 78, 98, 111]. In autonomous driving, video diffusion models [19,26,51,52,66,84,109] likewise target temporal coherence, but often add extra conditioning layers that increase memory and weaken pretrained priors.

### 2.2 Multi-view generation in autonomous driving

Advances in multi-view driving scene generation [31, 53, 80, 94, 97, 112, 114] have improved the realism and controllability of autonomous-driving simulators. For multi-view generation [12, 33, 42, 50], systems commonly inject structured cues, 3D bounding boxes, BEV maps, ego trajectories, and multi-camera poses, via ControlNet to enforce spatial and geometric constraints across views [14, 16, 38, 73, 85, 93, 96, 105]. BEV-based frameworks [49, 73, 90] enhance scene-level spatial coherence. Panacea [84] uses BEV and 4D spatiotemporal attention for coherence. MagicDrive [16] adopts tailored encoders and cross-view attention for precise 3D geometry. DriveArena [93] uses a reference frame as a conditioning signal to propagate information across sequences. Moreover, MagicDrive-V2 [15], Genesis [20] and DrivingSphere [89] use DiT architectures, enabling high-resolution 3D scene reconstruction. DiST-4D [19] leverages metric depth to align spatial structure and temporal continuity. These methods typically rely on extra training or parameters to enforce multi-view and temporal consistency.

### 2.3 Scene-level data augmentation

Scene-level data augmentation [1, 8, 40, 67, 87, 95], enhances data diversity by editing or transforming existing scenes, thereby complementing costly real-world data collection. One major approach is spatial manipulation [9,44,76], which involves reconstructing environments in 3D using NeRF [13, 45, 55, 100, 101] or Gaussian Splatting [34, 58, 92, 102, 108, 116], and then re-rendering them to preserve multi-view geometric consistency. Meanwhile, recent works [47, 86, 99, 113, 115] leverage diffusion-based scene editing to achieve geometry-aware manipulations. Another major approach is style-based augmentation [29, 30, 39], which modifies the visual domain without changing the scene geometry. These methods, implemented via neural style transfer, can be categorized into three strands:

(i) condition-specific transfer adjusts weather, lighting, and season to narrow out-of-distribution gaps [3, 43, 60, 64, 72, 107]; (ii) semantic-guided synthesis uses pixel-aligned masks or layouts for high-fidelity edits while preserving labels [81, 103, 106]; (iii) instruction-driven editing [5, 17, 23, 54] follows natural-language prompts, often via cross-attention control. Building on these advances, we focus on multi-view driving scene generation and propose a framework that synthesizes diverse, geometrically consistent driving scenarios from a text prompt.

## 3 Methods

### 3.1 Preliminary: latent diffusion models

Latent Diffusion Models (LDM) [63] train the diffusion process in a learned latent space. A perceptual autoencoder $(E, D)$ maps images to latents $\mathbf{z} = E(\mathbf{x})$ and reconstructs $\hat{\mathbf{x}} = D(\mathbf{z})$, reducing spatial complexity while preserving semantics. Diffusion is applied to $\mathbf{z}$, *i.e.*, $q(\mathbf{z}_t|\mathbf{z}_0)$ and $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$, with the noise-prediction loss. Conditioning (*e.g.*, text) is injected via cross-attention inside the U-Net backbone, enabling flexible, scalable generation with substantially reduced memory and compute compared to pixel-space diffusion. In FrozenDrive, we adopt the pretrained LDM, Stable Diffusion, for multi-view image generation, as it is lightweight, highly adaptable, and widely used [16, 84, 93, 94]. While others customize the SD, we use it frozen to leverage all the knowledge it contains.

### 3.2 Key insight: knowledge preservation

Pretrained LDMs, when used directly, are difficult to guide toward reliable generation of a realistic multi-view driving scene. A common remedy is to adapt the model with driving scene data—by (1) fine-tuning all or part of the backbone, (2) adding trainable adapters or modules, or (3) using a ControlNet-based conditioning pipeline [106]; these strategies are often combined. Our goal, however, is not only to synthesize plausible scenes but to retain the pretrained prior so that text-guided *zero-shot* synthesis remains effective.

Guided by this goal, we consider these strategies through the lens of knowledge preservation. It is widely recognized that full or partial fine-tuning [21, 35, 36, 46, 48, 56, 59, 110] can fit the training set while eroding the pretrained prior, weakening text alignment and zero-shot ability. More critically, we empirically observe that adding only a few trainable parameters (*e.g.*, multi-view or temporal cross-attention attached to the backbone) induces similar drift, narrowing visual diversity and weakening prompt fidelity (see Sec. 5.2). These observations motivate a simple principle: *freeze the backbone to preserve knowledge*. We therefore adopt ControlNet-based conditioning, where the LDM remains intact while ControlNet ingests structured signals and steers generation toward driving scenes. This design preserves the model's text understanding and enables robust zero-shot synthesis, while still allowing precise, task-specific control.

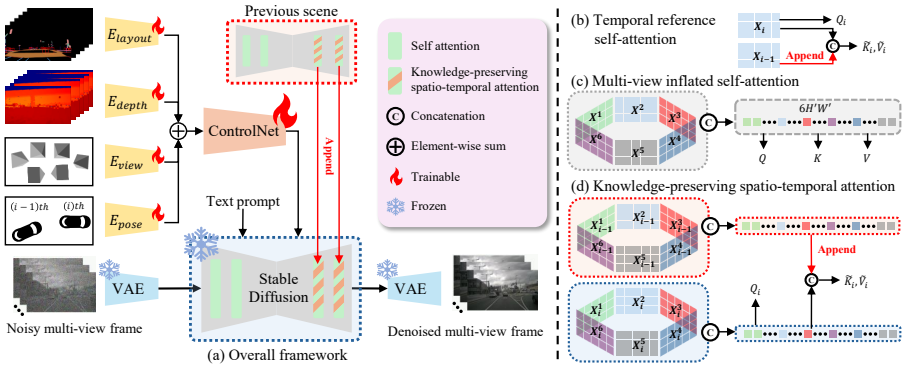



**Fig. 2: Overall framework.** (a) FrozenDrive uses five conditions: (i) an HD map, (ii) per-view camera indicator, (iii) depth map, (iv) relative pose, and (v) text description. The first four are multi-view, pixel- and spatially aligned and injected into the diffusion model via ControlNet, while the last enables zero-shot text guidance. After the encoder, we replace the original self-attention with our knowledge-preserving spatio-temporal attention. (b) Temporal reference self-attention enforces frame-to-frame consistency by expanding the input context to reuse the previous frame as a reference. (c) Multi-view inflated self-attention enforces cross-view consistency by concatenating latent features from all camera views. (d) Knowledge-preserving spatio-temporal attention reshapes self-attention so that multi-view inflated and temporal reference self-attention act jointly. During this process, the projection layers are kept frozen.

### 3.3 Overall pipeline

Figure 2 provides an overview of our pipeline. For synthesis, we use (i) an HD map with road-layout layers and 3D bounding boxes, (ii) per-view camera indicator, (iii) depth derived from an occupancy map, (iv) the relative pose to the previous frame, and (v) a text description. Each non-text signal $c_k$ is passed through a lightweight embedding network $E_k$ to produce an embedding $\mathbf{e}_k = E_k(c_k)$ in the latent space. These embeddings $\{\mathbf{e}_{\text{layout}}, \mathbf{e}_{\text{view}}, \mathbf{e}_{\text{depth}}, \mathbf{e}_{\text{pose}}\}$ are added to the latent feature map $\mathbf{z}$ and jointly serve as input to ControlNet. The resulting control features modulate the frozen diffusion backbone, while text is encoded separately by the frozen text encoder. Together, they guide generation without modifying or adding parameters to the diffusion backbone. Details of the embedding design are provided in the *supplementary materials*. To preserve prior knowledge, we keep this backbone frozen and promote multi-view/temporal consistency by restructuring self-attention context, *i.e.*, attended features(Sec. 3.4), integrating rich spatial cues while avoiding pretrained knowledge forgetting. Finally, because fidelity tends to track observation frequency, the long-tail distribution impairs rare-class rendering, and we mitigate this with an object-focused ratio loss that upweights underrepresented classes (Sec. 3.5).

### 3.4 Knowledge-preserving multi-view generation

Our objective is to synthesize realistic autonomous driving scenes across multiple cameras and over time while preserving the prior of a pretrained diffusion model. We therefore adopt a ControlNet-based pipeline. Yet directly conditioning a frozen generator is insufficient for our setting: multi-camera perception demands **cross-view consistency** (overlapping views must agree on geometry, appearance, and semantics), and video rollout requires **temporal consistency** (successive frames must remain coherent under ego- and object motion without discontinuities). Enforcing both under a knowledge-preserving regime is challenging, as we introduce no new trainable parameters into the diffusion backbone.

To address this, we propose a *knowledge-preserving spatio-temporal attention* that operates on the frozen LDM. Rather than learning new attention modules or modifying existing parameters, we reinterpret how the original self-attention layers are *fed* by manipulating their inputs. The block is composed of two complementary mechanisms: (i) *multi-view inflated self-attention*, which promotes cross-view agreement by allowing features from different camera views to attend to one another and (ii) *temporal reference self-attention*, which encourages frame-to-frame stability by letting the current frame attend to the previous frame as an explicit reference. Both mechanisms reshape the attention context seen by each layer while leveraging conditional signals injected by ControlNet, leaving the attention weights themselves untouched. We next describe the attention module and its auxiliary embeddings.

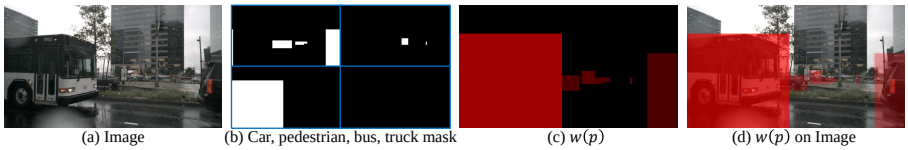
**Multi-view inflated self-attention.** To enforce cross-view consistency while preserving the pretrained prior, we introduce multi-view inflated self-attention (MISA). MISA concatenates latent features from all $n_{\text{view}}$ cameras and performs a single self-attention pass with the frozen layer, enabling cross-view exchange beyond standard within-view attention. ControlNet supplies per-view conditioning features, while MISA expands the attention field across views. Let $\mathbf{X}^{(v)} \in \mathbb{R}^{N_v \times d_{\text{model}}}$ denote the latent feature token for view $v$, where $N_v$ is the number of tokens for view $v$ and $d_{\text{model}}$ is the token embedding width. We form a joint sequence and apply shared self-attention:

$$\tilde{\mathbf{X}} := \left[\, \mathbf{X}^{(1)} \,;\, \mathbf{X}^{(2)} \,;\, \cdots \,;\, \mathbf{X}^{(n_{\text{view}})} \,\right], \tag{1}$$

$$\text{Attn}(\tilde{\mathbf{X}}) = \text{softmax}\left(\frac{(\tilde{\mathbf{X}}\mathbf{W}^{Q})(\tilde{\mathbf{X}}\mathbf{W}^{K})^{\top}}{\sqrt{d_h}}\right)(\tilde{\mathbf{X}}\mathbf{W}^{V}), \tag{2}$$

where $\mathbf{W}^{Q}, \mathbf{W}^{K}, \mathbf{W}^{V} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are frozen LDM parameters and $d_h$ is the per-head dimension.

We also indicate which view produced each feature via a lightweight view embedding $\mathbf{e}_{\text{view}}$. Each view index $v \in \{1, \ldots, n_{\text{view}}\}$ is encoded with Fourier features [74] and supplied to ControlNet as an additional conditioning input. This simple tag provides view identity and neighborhood structure, improving cross-view correspondence without introducing any new attention parameters.



(a) Image (b) Car, pedestrian, bus, truck mask (c) $w(p)$ (d) $w(p)$ on Image

**Fig. 3: Details of the class-wise weighting mask.** (a) Input image. (b) Binary masks for car, pedestrian, bus, and truck. (c) Weighted mask $w(p)$ obtained by scaling each category by its object-presence ratio. In overlapping regions, we apply a max rule and keep the larger value. (d) Overlay of $w(p)$ on the image.

**Temporal reference self-attention.** To enforce frame-to-frame consistency under knowledge preservation, we introduce temporal reference self-attention (TRSA), which, analogously to MISA, keeps all LDM attention weights frozen and instead expands their input context to reuse the previous frame as a reference. Concretely, for the current frame $i$ and a reference frame $i-j$, we take their latent features at the same noise level, $\mathbf{X}_i$ and $\mathbf{X}_{i-j}$, and compute the projections using frozen weights. We obtain the query, key, and value for the current frame as $\mathbf{Q}_i = \mathbf{X}_i\mathbf{W}^Q$, $\mathbf{K}_i = \mathbf{X}_i\mathbf{W}^K$, and $\mathbf{V}_i = \mathbf{X}_i\mathbf{W}^V$, and the key and value for the reference frame as $\mathbf{K}_{i-j} = \mathbf{X}_{i-j}\mathbf{W}^K$ and $\mathbf{V}_{i-j} = \mathbf{X}_{i-j}\mathbf{W}^V$.

We then augment the key and value banks of the current frame with the reference, and the self-attention is computed as :

$$\tilde{\mathbf{K}}_i = \big[\mathbf{K}_i\,;\,\mathbf{K}_{i-j}\big], \quad \tilde{\mathbf{V}}_i = \big[\mathbf{V}_i\,;\,\mathbf{V}_{i-j}\big], \tag{3}$$

$$\mathrm{Attn}(\mathbf{X}_i) = \mathrm{softmax}\left(\frac{\mathbf{Q}_i\tilde{\mathbf{K}}_i^\top}{\sqrt{d_h}}\right)\tilde{\mathbf{V}}_i. \tag{4}$$

Exposing each current features to keys and values from the previous frame enables the retrieval of temporally aligned appearance and geometry cues (*e.g.*, object, weather), stabilizing the rollout without introducing new parameters.

We also encode inter-frame motion with a lightweight relative pose embedding $\mathbf{e}_{\text{pose}}$ between frames $i-j$ and $i$. Instead of passing the raw 6-DoF transform, we construct a spatial map: for each image location $(x, y)$, lift to $(x, y, 0)$, transform by the known 3D relative pose, take the in-plane coordinates $(x', y')$, and encode them with Fourier features [74]. This pose-aware map is injected via ControlNet as additional conditioning, providing dense correspondence cues.

**Knowledge-preserving spatio-temporal attention.** Finally, we unify the above mechanisms by reshaping self-attention so that MISA and TRSA act jointly within a single, frozen block. For effective interaction with ControlNet, we apply this spatio-temporal attention only in the LDM decoder: ControlNet supplies scene-level conditioning (*e.g.*, HD maps, depth, view, and relative poses), and the decoder fuses these signals with our attention to produce view-consistent, temporally stable generations. This design keeps the diffusion backbone frozen and parameter-free, thereby preserving the model's generalization capability.

**Table 1: Generation fidelity experiments.** We generate multi-view images under nuScenes [6] validation set conditions. Metrics are computed by evaluating the synthesized results with UniAD [27] and measuring FVD [77]. **Bold** and underline denote the best and second-best, respectively.

| Method | Venue | Backbone | 3DOD (↑) | | BEV Segmentation mIoU (↑) | | | | L2 (↓) | | | | Quality (↓) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | mAP | NDS | Lanes | Drivable | Divider | Crossing | 1.0s | 2.0s | 3.0s | Avg. | FVD |
| nuScenes [6] | - | - | 37.98 | 49.85 | 31.31 | 69.14 | 25.93 | 14.36 | 0.51 | 0.98 | 1.65 | 1.05 | - |
| MagicDrive [16] | ICLR'24 | SD [63] | 12.92 | 28.36 | 21.95 | 51.46 | 17.10 | 5.25 | 0.57 | 1.14 | 1.95 | 1.22 | 218.1 |
| Panacea [84] | CVPR'24 | SD [63] | 13.72 | 27.73 | 18.23 | 52.37 | 17.21 | - | 0.58 | 1.14 | 1.97 | 1.23 | 139.0 |
| DrivingSphere [89] | CVPR'25 | STDiT [111] | 21.45 | 34.16 | 27.99 | 62.87 | 22.29 | - | 0.54 | 1.10 | 1.76 | 1.13 | 103.4 |
| DiST-4D [19] | ICCV'25 | STDiT [111] | 15.63 | 32.44 | 26.80 | 60.32 | 21.69 | 10.99 | 0.56 | 1.11 | 1.91 | 1.19 | **22.6** |
| DriveArena [93] | ICCV'25 | SD [63] | 16.06 | 30.03 | 26.14 | 59.37 | 20.79 | 8.92 | 0.56 | 1.10 | 1.89 | 1.18 | 185.3 |
| MagicDrive-V2 [15] | ICCV'25 | STDiT [111] | 15.24 | 31.25 | 25.02 | 58.63 | 19.84 | 9.98 | **0.49** | 1.00 | 1.78 | 1.09 | 81.6 |
| X-Scene [94] | NeurIPS'25 | SD [63] | 20.40 | 31.76 | 28.04 | 61.96 | 22.32 | 10.48 | 0.55 | 1.08 | 1.81 | 1.15 | 179.7 |
| **FrozenDrive (Ours)** | - | SD [63] | **21.87** | **35.32** | **28.88** | **64.27** | **23.58** | **11.66** | 0.50 | **0.98** | **1.66** | **1.05** | 136.8 |

### 3.5 Objective function

Guided by our design, we train ControlNet and the embedders, keeping the diffusion backbone frozen, to predict noise with the standard DDPM [24] objective:

$$\mathcal{L} = \mathbb{E}_{x_0,c,\epsilon,t}\big[\big\|\epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon,\ t,\ c\big)\big\|^2\big]. \tag{5}$$

Supervised driving data are often long-tailed (*e.g.*, in the nuScenes dataset, *bicycle* appears at only ∼2.3% of the *car* frequency; see Table A in *supplementary materials*), which hurts fidelity for rare classes. To mitigate this, we propose object-presence ratio loss, an object-balanced ratio loss that emphasizes underrepresented categories via a per-pixel weight map derived from projected 3D boxes (see Fig. 3):

$$\mathcal{L}_{\text{total}} = \frac{1}{|\Omega|}\sum_{p\in\Omega}\big(1+\lambda\, w(p)\big)\,\mathcal{L}(p), \tag{6}$$

$$w(p) = \max_{k\in\mathcal{K}_p} w_k, \quad \mathcal{K}_p = \{\, k \mid p \in \mathcal{M}_k \,\}. \tag{7}$$

Here, $\Omega$ is the image lattice and $\mathcal{L}(p)$ the per-pixel diffusion loss. $\mathcal{M}_k$ denotes the foreground mask obtained by projecting class-$k$ 3D boxes and $w_k$ is a class-dependent weight (larger for rarer classes). When there is no object mask at pixel $p$, *i.e.*, $\mathcal{K}_p = \varnothing$, set $w(p) = 0$. The max-overlap rule lets the rarest overlapping instance dominate the weighting.

## 4 Experiments

### 4.1 Experimental settings

We train and evaluate our proposed method on the nuScenes [6] dataset. It provides rich sensor data and annotations, including LiDAR point clouds, multi-view images, lane information, poses, and related labels. Similar to prior work [16,82], the 2Hz keyframe annotations are temporally interpolated to obtain 12Hz labels. For the depth map, we stack LiDAR point clouds to generate occupancy


**Fig. 4: Qualitative comparison of generated samples** under nuScenes [6] validation conditions. Using the same nuScenes conditions, we generate multi-view images with each method to evaluate visual quality. Compared to the ground truth (top row), MagicDrive-V2 [15] produces structural artifacts in road conditions (red boxes), and DriveArena [93] exhibits view-inconsistent object shapes (blue boxes). Our method better preserves scene layout while maintaining superior multi-view consistency.

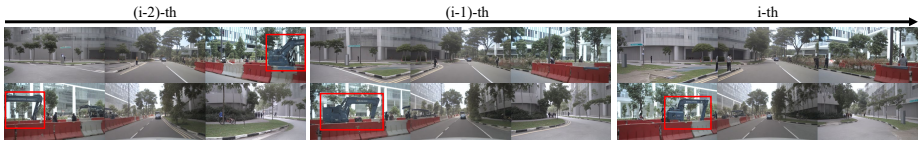



**Fig. 5: Visualization of sequential generation.** FrozenDrive produces high-fidelity, temporally coherent images across scenes. The rare construction vehicle (red boxes) is faithfully preserved in shape, color, and position across viewpoints and consecutive frames, demonstrating strong object-level consistency.

and measure the depth to the ego position [37]. We first assess the FrozenDrive generation quality in Sec. 4.2. In Sec. 4.3, we further evaluate the impact of FrozenDrive-based data augmentation on adverse conditions and measure how these samples improve the generalization of an existing AD model.

### 4.2 Generation fidelity

**Metric.** Following previous studies [16, 89, 93], we compare the performance of FrozenDrive with other methods in terms of perception and planning via UniAD [27]. For generation quality, we utilize FVD [77] to evaluate.
**Baselines** are MagicDrive [16], Panacea [84], DrivingSphere [89], DiST-4D [19], DriveArena [93], MagicDrive-V2 [15], and X-Scene [94] which generates multi-view images under nuScenes dataset conditions.
**Results.** In Tab. 1, our method achieves the strongest overall downstream performance among generative approaches, attaining the best 3D detection mAP/ NDS and BEV segmentation mIoU. The particularly strong BEV segmentation

**Table 2: Impact of Synthetic Night/Rain Data on nuScenes Perception and Planning.** Perception and planning performance of SparseDrive [71] on nuScenes under adverse conditions (night and rain) with different data generation strategies. The baseline uses only original normal-weather samples, while rule-based methods [41, 75] translate normal samples into night/rain images with hand-crafted filters that introduce artifacts. DriveArena [93], MagicDrive-V2 [15], and FrozenDrive synthesize condition-aware data from text prompts. **Bold** and underline denote the best and second-best.

| | Method | 3DOD (↑) | | Online mapping results (↑) | | | | L2(m) (↓) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | NDS | $AP_{ped}$ | $AP_{divider}$ | $AP_{boundary}$ | mAP | 1.0s | 2.0s | 3.0s | Avg. |
| Night | Baseline | 6.62 | 15.19 | 0.07 | 8.48 | 9.42 | 5.99 | 0.74 | 1.36 | 2.11 | 1.40 |
| | Rule-based [75] | 7.42 | 13.11 | 0.30 | 8.93 | 10.27 | 6.50 | 0.63 | 1.20 | 1.90 | 1.24 |
| | DriveArena [93] | 8.89 | 17.52 | 0.29 | 9.78 | 10.93 | 7.00 | 0.54 | 1.04 | 1.68 | 1.09 |
| | MagicDrive-V2 [15] | 12.68 | 17.32 | 1.35 | 14.37 | 19.36 | 11.69 | 0.54 | 1.07 | 1.73 | 1.11 |
| | **FrozenDrive (Ours)** | **18.15** | **24.95** | **6.54** | **22.29** | **34.25** | **21.03** | **0.44** | **0.89** | **1.47** | **0.93** |
| Rain | Baseline | 31.60 | 41.20 | 25.50 | 25.47 | 23.28 | 24.75 | 0.35 | 0.70 | 1.18 | 0.75 |
| | Rule-based [41] | 30.76 | 39.90 | 27.06 | 24.96 | 23.58 | 25.20 | 0.36 | 0.72 | 1.20 | 0.76 |
| | DriveArena [93] | 33.46 | **45.25** | 29.19 | 32.66 | 25.32 | 29.06 | 0.33 | 0.65 | 1.08 | 0.71 |
| | MagicDrive-V2 [15] | 33.93 | 42.20 | 31.96 | 32.72 | 25.39 | 30.02 | 0.34 | 0.69 | 1.15 | 0.73 |
| | **FrozenDrive (Ours)** | **35.15** | 44.26 | **33.13** | **33.00** | **28.04** | **31.39** | **0.29** | **0.55** | **0.90** | **0.58** |

results indicate that our generated scenes exhibit high multi-view spatial consistency. We further assess generation fidelity with FVD: among SD-based image diffusion baselines, FrozenDrive achieves the lowest FVD, while STDiT-based video diffusion models still obtain lower FVD thanks to explicit temporal modeling. At the same time, planning models tested on our generated data achieve the best metrics, closely matching the performance on real nuScenes data. Beyond superior quantitative results, our model also produces higher-quality fine-grained details. As shown in Fig. 4, the competing method [93] often removes scene elements or produces view-inconsistent object shapes, whereas our method better preserves the original scene and maintains strong multi-view consistency. Additionally, we visualize our sequential generation results through Fig. 5. Even in complex driving scenarios, our method generates highly competitive, temporally consistent images without relying on heavy video diffusion models.

### 4.3 Data augmentation for autonomous driving

We apply our domain-targeted augmentation with FrozenDrive, driven solely by text prompts, to synthesize specific adverse conditions while preserving the original annotations. We evaluate on the nuScenes benchmark, focusing on the *rain* and *night* conditions. Based on the official metadata, we partition the dataset into three disjoint subsets: night, rain, and normal. For each target domain $r \in \{\text{night}, \text{rain}\}$, we construct an augmented training set: given a normal sample $x$, we synthesize its adverse-weather counterpart $x^r$ with probability $p$.

**Metric.** We evaluate FrozenDrive against baselines by training SparseDrive [71] on each augmented dataset. We report perception and planning performance under corresponding adverse conditions to quantify the improvements.

**Baseline.** We compare FrozenDrive-based augmentation against three competitors: Baseline, which uses only original normal-weather samples, Rule-based methods [41, 75], which translate normal samples into night/rain images with

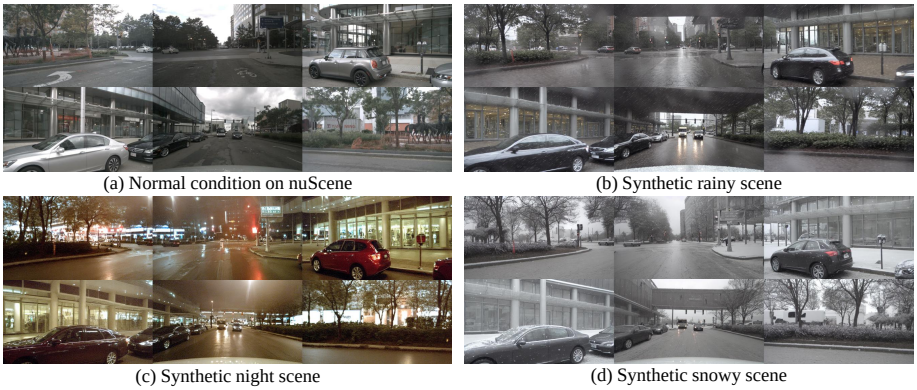



**Fig. 6: Examples of text-driven weather-augmented samples.** Given text prompts specifying the target weather, conditioned on (a) normal scenes, FrozenDrive generates (b) synthetic rain, (c) synthetic night, and (d) synthetic snowy scenes. In the fifth view, raindrops on the car hood (rain) and strong light sources with their reflections on the road (night) are clearly generated.

hand-crafted filters, and DriveArena [93], which, like FrozenDrive, a Stable Diffusion-based text-to-image synthesis framework, and MagicDrive-V2 [15], a diffusion transformer (STDiT)-based model for text-conditioned data generation.
**Results.** Table 2 reports the effect of data augmentation on perception and planning. Under night conditions, the AD model trained with FrozenDrive-augmented data attains the best overall performance, surpassing all baselines with higher detection and mapping scores and significantly lower planning errors. A similar trend is observed under rainy conditions. These results indicate that the scene generation model trained with our knowledge-preserving scheme acquires stronger generative capability by text-prompting for rare adverse scenarios, which in turn leads to better autonomous driving performance under such conditions. Moreover, as illustrated in Fig. 6, our method can faithfully synthesize a broad range of weather conditions, including previously unseen scenarios such as snowy scenes. Further details are provided in the *supplementary materials.*

## 5 Analysis

### 5.1 Ablation studies

**Knowledge-preserving spatio-temporal attention.** We conduct toy ablations on our knowledge-preserving spatio-temporal attention (MISA and TRSA). As shown in Tab. 3, removing TRSA degrades temporal quality and worsens FVD while keeping BEV segmentation relatively high, whereas using only TRSA improves FVD but clearly harms BEV segmentation. The full model with both modules achieves the highest BEV mIoU and lowest FVD, confirming that MISA and TRSA are complementary for multi-view temporal continuity.

**Object-presence ratio loss.** We compare 3D object detection performance with and without the proposed object-presence ratio loss and break down the results by category. As shown in Tab. 4, object-presence ratio loss encourages more faithful object synthesis and yields consistent 3DOD gains, with particularly large improvements for rare categories such as motorcycle and bicycle (1.07% and 1.00% of the training set). This mitigates data imbalance in downstream detection. Additional per-category results and occurrence ratios are provided in the *supplementary metarials.*

### 5.2 Knowledge forgetting

Prior studies have reported knowledge forgetting in personalization and fine-tuning of pretrained models [2, 21, 36, 48, 57, 62, 65, 110]. In particular, they indicate that fine-tuning can induce drift in the model's text alignment [59]. Extending this observation to multi-view driving scene generation, we find that not only full/partial fine-tuning but also adding and training extra layers on top of a frozen backbone can induce similar drift.

**Qualitative comparison.** To isolate this effect, we construct a controlled baseline by replacing our parameter-free MISA and TRSA with learnable multi-view cross-attention and temporal cross-attention layers attached to the pretrained backbone (denoted as "SD + CA".) Figure 7 (a) shows generation results with text prompt, "*Snowy weather. Heavy snow.*" While both our method and "SD + CA" successfully ensure multi-view and temporal consistency, "SD + CA" tends to degrade alignment and zero-shot controllability, indicating a forgetting of the pretrained prior. In contrast, our knowledge-preserving spatio-temporal attention enforces multi-view and temporal consistencies solely through input reshaping, while preserving text alignment.

**Clip score.** This behavior is further supported by the CLIP scores in Tab. 5, which we employ as a relative measure to evaluate stylistic alignment against the reference samples. We compare our method against DriveArena [93] and MagicDrive-V2 [15], while treating normal-weather nuScenes samples and real-weather data as references. For seen conditions such as night and rain, DriveArena [93] and MagicDrive-V2 [15] achieve broadly comparable scores, suggesting that these models effectively mimic global weather textures. However, while these CLIP scores reflect stylistic "plausibility," our superior downstream performance in perception and planning (Tab. 2) provides more definitive evidence that our method not only preserves essential scene knowledge but also excels in the detailed realization of complex scene elements, such as precise road layouts and object instances. This distinction becomes even more pronounced in the snow setting, which is entirely absent from the training data. In this unseen scenario, our method surpasses existing generative baselines by a substantial margin, highlighting that our knowledge-preserving design is uniquely capable of generalizing to novel out-of-distribution conditions.

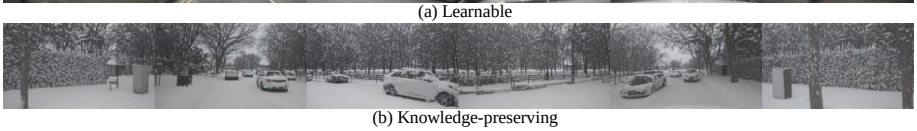
**Table 3: Ablation of knowledge-preserving spatio-temporal attention.** We evaluate BEV segmentation for multi-view consistency and FVD for temporal continuity.

| Method | | BEV Segmentation mIoU (%) (↑) | | | | Quality (↓) |
|---|---|---|---|---|---|---|
| MISA | TRSA | Lanes | Drivable | Divider | Crossing | FVD |
| ✓ | ✓ | **25.6** | **56.5** | **20.9** | **8.9** | **144.1** |
| ✓ | | 25.1 (-0.5) | 55.1 (-1.4) | 20.5 (-0.4) | 7.8 (-1.1) | 174.2 (+**30.1**) |
| | ✓ | 23.4 (**-2.2**) | 51.9 (**-4.6**) | 18.6 (**-2.3**) | 5.6 (**-3.3**) | 145.1 (+1.0) |

**Table 4: Ablation of object-presence ratio loss.** We evaluate 3DOD metric w/ and w/o the proposed loss, broken down by category. Categories are ordered left to right by decreasing frequency.

| Method | 3DOD (↑) | | Object category (AP, ↑) | | | |
|---|---|---|---|---|---|---|
| | mAP | NDS | Car | Pedestrian | Motorcycle | Bicycle |
| w/ loss | **21.9** | **35.3** | **38.6** | **29.5** | **10.0** | **9.2** |
| w/o loss | 16.1 (-5.8) | 29.6 (-5.7) | 33.5 (-5.1) | 25.3 (-4.2) | 0.4 (-9.6) | 1.9 (-7.3) |

(a) Learnable



(b) Knowledge-preserving

**Fig. 7: Zero-shot text-guided generation results.** Examples from a model with (a) learnable multi-view/temporal cross-attention and (b) our parameter-free knowledge-preserving spatio-temporal attentions with a “Snowy weather" prompt.

### 5.3 Consistency evaluation

While the previous ablation study provides indirect evidence of the consistency improvements brought by MISA and TRSA through BEV segmentation performance, we further investigate this aspect using two primary metrics that directly assess consistency. This complementary analysis allows us to examine the effect of MISA and TRSA on consistency more explicitly.

**Lane alignment.** To evaluate geometric consistency, we measure the degree of lane alignment, defined as the mIoU between the ground-truth and generated lanes in the image space. As shown in Tab. 6, our FrozenDrive provides geometric consistency comparable to the baseline that requires additional learnable cross-attention layers with partial fine-tuning (the same as "SD + CA" of Sec. 5.2). This indicates that our method achieves high geometric accuracy without introducing additional parameters to the diffusion backbone.

**Temporal consistency.** To assess temporal consistency, we adopt the subject and background consistency metrics from VBench [28]. Subject consistency is calculated based on DINO [7] feature similarities across frames to capture object-wise semantic correspondence, while background consistency utilizes CLIP [61] similarities to evaluate global stability. Compared to MagicDrive-V2 [15], our method shows slightly lower consistency scores, which is expected since MagicDrive-V2 is built on a video diffusion model. Nevertheless, our FrozenDrive achieves stronger temporal consistency than the SD + CA and DriveArena [93], both of which are fine-tuned baselines.

**Table 5:** CLIP scores for weather-conditioned text–image generation.

| Method | Rain | Night | Snow |
|---|---|---|---|
| Normal weather | 0.2304 | 0.2319 | 0.1950 |
| Real target weather | 0.2564 | 0.2662 | 0.2711 |
| DriveArena [93] | 0.2570 | 0.2716 | 0.2277 |
| MagicDrive-V2 [15] | 0.2616 | 0.2705 | 0.2181 |
| FrozenDrive (Ours) | 0.2595 | 0.2751 | 0.2507 |

**Table 6:** Lane alignment and Vbench [28] temporal consistency evaluation.

| Method | Lane (↑) | Vbench (↑) | |
|---|---|---|---|
| | mIoU | Subject | BG |
| SD + CA | <u>33.55</u> | 0.7586 | 0.8679 |
| DriveArena [93] | 30.73 | 0.7573 | 0.8556 |
| MagicDrive-V2 [15] | 26.71 | **0.7966** | **0.8845** |
| FrozenDrive (Ours) | **33.58** | <u>0.7724</u> | <u>0.8700</u> |

### 5.4 Conclusion

We propose FrozenDrive, a zero-shot text-guided multi-view driving scene generator that steers a parameter-free frozen pretrained diffusion model. With our knowledge-preserving attention, FrozenDrive produces realistic, spatio-temporally coherent multi-view driving scenes via driving-stack signals and text. Moreover, FrozenDrive-augmented data improves downstream perception and planning, especially under rare and adverse conditions.

**Limitations and future works.** Our parameter-free frozen diffusion model offers good text alignment and controllability, but still lags behind recent video diffusion models in long-range temporal coherence. Moreover, existing protocols do not directly evaluate data augmentation quality, and developing more faithful metrics remains an important direction for future work. Extending our knowledge-preserving design beyond Stable Diffusion to stronger video diffusion or DiT-based backbones is another promising direction.

## Acknowledgements

This work was supported by the Technology Innovation Program (2410018126, KT224355, Development of autonomous driving connectivity technology based on sensor-infrastructure cooperation) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea), and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25473963). This work was also funded by 42dot Inc. We appreciate the support and valuable discussions from the members of the 42dot research team.

## References

1. Abu Alhaija, H., Mustikovela, S.K., Mescheder, L., Geiger, A., Rother, C.: Augmented reality meets computer vision: Efficient data generation for urban driving scenes. International Journal of Computer Vision **126**(9), 961–972 (2018)
2. Aghajanyan, A., Shrivastava, A., Gupta, A., Goyal, N., Zettlemoyer, L., Gupta, S.: Better fine-tuning by reducing representational collapse. arXiv preprint arXiv:2008.03156 (2020)
3. Assion, F., Gressner, F., Augustine, N., Klemenc, J., Hammam, A., Krattinger, A., Trittenbach, H., Philippsen, A., Riemer, S.: A-bdd: Leveraging data augmentations for safe autonomous driving in adverse weather and lighting. arXiv preprint arXiv:2408.06071 (2024)

4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
8. Chen, Y., Rong, F., Duggal, S., Wang, S., Yan, X., Manivasagam, S., Xue, S., Yumer, E., 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. pp. 7230–7240 (2021)
9. Cheng, H., Xu, J., Peng, L., Yang, Z., He, X., Wu, B.: Object-level data augmentation for visual 3d object detection in autonomous driving. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
10. Cho, H., Kang, J.Y., Lee, G., Yang, H., Park, H., Jung, S., Yoon, K.J.: Vr-drive: Viewpoint-robust end-to-end driving with feed-forward 3d gaussian splatting. Advances in Neural Information Processing Systems **38**, 114830–114851 (2026)
11. Cho, H., Lee, G., Kang, J.Y., Yang, H., Park, H., Yoon, K.J.: Heat: Heterogeneous end-to-end autonomous driving via trajectory-guided world models. arXiv preprint arXiv:2605.19631 (2026)
12. Dong, H., Wang, X., Lin, D., Wu, Y., Chen, Q., Liu, R., Yang, K., Li, P., Guo, Q.: Noisecontroller: Towards consistent multi-view video generation via noise decomposition and collaboration. arXiv preprint arXiv:2504.18448 (2025)
13. Feldmann, C., Siegenheim, N., Hars, N., Rabuzin, L., Ertugrul, M., Wolfart, L., Pollefeys, M., Bauer, Z., Oswald, M.R.: Nerfmentation: Nerf-based augmentation for monocular depth estimation. arXiv preprint arXiv:2401.03771 (2024)
14. Gao, R., Chen, K., Li, Z., Hong, L., Li, Z., Xu, Q.: Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. arXiv preprint arXiv:2405.14475 (2024)
15. Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28135–28144 (2025)
16. Gao, R., Chen, K., Xie, E., HONG, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=sBQwvucduK
17. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
18. Gu, J., Hu, C., Zhang, T., Chen, X., Wang, Y., Wang, Y., Zhao, H.: Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5496–5506 (2023)
19. 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)
20. Guo, X., Wu, Z., Xiong, K., Xu, Z., Zhou, L., Xu, G., Xu, S., Sun, H., WANG, B., Chen, G., Ye, H., Liu, W., Wang, X.: Genesis: Multimodal driving scene generation with spatio-temporal and cross-modal consistency. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=Q7YnqREWLq
21. He, T., Liu, J., Cho, K., Ott, M., Liu, B., Glass, J., Peng, F.: Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1121–1133 (2021)
22. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
23. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=_CDixzkzeyb
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
25. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
26. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
27. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)
28. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
29. Islam, K., Zaheer, M.Z., Mahmood, A., Nandakumar, K.: Diffusemix: Label-preserving data augmentation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27621–27630 (2024)
30. Jeong, Y., Cho, H., Yoon, K.J.: Towards robust event-based networks for nighttime via unpaired day-to-night event translation. In: European Conference on Computer Vision. pp. 286–306. Springer (2024)
31. Ji, Y., Zhu, Z., Zhu, Z., Xiong, K., Lu, M., Li, Z., Zhou, L., Sun, H., Wang, B., Lu, T.: Cogen: 3d consistent video generation via adaptive conditioning for autonomous driving. arXiv preprint arXiv:2503.22231 (2025)
32. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)

33. Jiang, J., Hong, G., Zhang, M., Hu, H., Zhan, K., Shao, R., Nie, L.: Dive: Efficient multi-view driving scenes generation based on video diffusion transformer. arXiv preprint arXiv:2504.19614 (2025)
34. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
35. Kotha, S., Springer, J., Raghunathan, A.: Understanding catastrophic forgetting in language models via implicit inference. In: NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models (2024), https://openreview.net/forum?id=wkQy8mLIb9
36. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)
37. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11971–11981 (2025)
38. Li, H., Yang, Z., Qian, Z., Zhao, G., Huang, Y., Yu, J., Zhou, H., Liu, L.: Dualdiff: Dual-branch diffusion model for autonomous driving with semantic fusion. In: IEEE International Conference on Robotics and Automation (ICRA). IEEE (2025), https://arxiv.org/abs/2505.01857
39. Li, J., Li, B., Tu, Z., Liu, X., Guo, Q., Juefei-Xu, F., Xu, R., Yu, H.: Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15205–15215 (2024)
40. Li, N., Song, F., Zhang, Y., Liang, P., Cheng, E.: Traffic context aware data augmentation for rare object detection in autonomous driving. In: 2022 international conference on robotics and automation (ICRA). pp. 4548–4554. IEEE (2022)
41. Li, R., Cheong, L.F., Tan, R.T.: Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1633–1642 (2019)
42. Li, X., Zhang, Y., Ye, X.: Drivingdiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model. In: European Conference on Computer Vision. pp. 469–485. Springer (2024)
43. Li, X., Kou, K., Zhao, B.: Weather gan: Multi-domain weather translation using generative adversarial networks. arXiv preprint arXiv:2103.05422 (2021)
44. Li, Y., Lin, Z.H., Forsyth, D., Huang, J.B., Wang, S.: Climatenerf: Extreme weather synthesis in neural radiance field. In: Proceedings of the ieee/cvf international conference on computer vision. pp. 3227–3238 (2023)
45. Li, Z., Li, L., Zhu, J.: Read: Large-scale neural scene rendering for autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1522–1529 (2023)
46. Li, Z., Ren, K., Jiang, X., Li, B., Zhang, H., Li, D.: Domain generalization using pretrained models without fine-tuning. arXiv preprint arXiv:2203.04600 (2022)
47. Liang, Y., Yan, Z., Chen, L., Zhou, J., Yan, L., Zhong, S., 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. vol. 39, pp. 5164–5172 (2025)
48. Liao, Y.C., Chen, J.J., Huang, C.P., Lin, C.S., Wu, M.L., Wang, Y.C.F.: Continual personalization for diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15511–15520 (2025)

49. Liu, B., Wang, K., Liu, Y., Bao, J., Han, T., Yu, J.: Mvpbev: Multi-view perspective image generation from bev with test-time controllability and generalizability. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8393–8401 (2024)
50. Lu, H., Wu, X., Wang, S., Qin, X., Zhang, X., Han, J., Zuo, W., Tao, J.: Seeing beyond views: Multi-view driving scene video generation with holistic attention. arXiv preprint arXiv:2412.03520 (2024)
51. Lu, J., Huang, Z., Yang, Z., Zhang, J., Zhang, L.: Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In: European Conference on Computer Vision. pp. 329–345. Springer (2024)
52. Lu, Y., Ren, X., Yang, J., Shen, T., Wu, Z., Gao, J., Wang, Y., Chen, S., Chen, M., Fidler, S., et al.: Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27272–27283 (2025)
53. Mei, J., Hu, T., Yang, X., Wen, L., Yang, Y., Wei, T., Ma, Y., Dou, M., Shi, B., Liu, Y.: Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. arXiv preprint arXiv:2409.04003 (2024)
54. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
55. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
56. Ming, Y., Li, Y.: How does fine-tuning impact out-of-distribution detection for vision-language models? International Journal of Computer Vision **132**(2), 596–609 (2024)
57. Mukhoti, J., Gal, Y., Torr, P.H., Dokania, P.K.: Fine-tuning can cripple your foundation model; preserving features may be the solution. arXiv preprint arXiv:2308.13320 (2023)
58. Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Huang, G., Liu, C., Chen, Y., Wang, Y., Zhang, X., et al.: Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1559–1569 (2025)
59. Park, N., Kim, K., Shim, H.: Textboost: Towards one-shot personalization of text-to-image models via fine-tuning text encoder. arXiv preprint arXiv:2409.08248 (2024)
60. Qian, C., Guo, Y., Mo, Y., Li, W.: Weatherdg: Llm-assisted procedural weather generation for domain-generalized semantic segmentation. IEEE Robotics and Automation Letters **10**(6), 5919–5926 (2025). https://doi.org/10.1109/LRA.2025.3559821
61. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
62. Ram, S., Neiman, T., Feng, Q., Stuart, A., Tran, S., Chilimbi, T.: Dreamblend: Advancing personalized fine-tuning of text-to-image diffusion models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3614–3623. IEEE (2025)
63. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

64. Rothmeier, T., Huber, W., Knoll, A.C.: Time to shine: Fine-tuning object detection models with synthetic adverse weather images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4447–4456 (2024)
65. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
66. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
67. Ryu, K., Hwang, S., Park, J.: Instant domain augmentation for lidar semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9350–9360 (2023)
68. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (October 2020), https://arxiv.org/abs/2010.02502
69. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=PxTIG12RRHS
70. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
71. Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S.: Sparsedrive: End-to-end autonomous driving via sparse scene representation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8795–8801. IEEE (2025)
72. Suryanto, N., Adiputra, A.A., Kadiptya, A.Y., Le, T.T.H., Pratama, D., Kim, Y., Kim, H.: Cityscape-adverse: Benchmarking robustness of semantic segmentation with realistic scene modifications via diffusion-based image editing. IEEE Access (2025). https://doi.org/10.1109/ACCESS.2025.3537981
73. Swerdlow, A., Xu, R., Zhou, B.: Street-view image generation from a bird's-eye view layout. IEEE Robotics and Automation Letters **9**(4), 3578–3585 (2024)
74. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020)
75. Thompson, W.B., Shirley, P., Ferwerda, J.A.: A spatial post-processing algorithm for images of night scenes. Journal of Graphics Tools **7**(1), 1–12 (2002)
76. Tong, W., Xie, J., Li, T., Li, Y., Deng, H., Dai, B., Lu, L., Zhao, H., Yan, J., Li, H.: 3d data augmentation for driving scenes on camera. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 46–63. Springer (2024)
77. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
78. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. CoRR (2025)
79. Wang, J., Yao, Y., Feng, X., Wu, H., Wang, Y., Huang, Q., Ma, Y., Zhu, X.: Stage: A stream-centric generative world model for long-horizon driving-scene simulation. arXiv preprint arXiv:2506.13138 (2025)

80. Wang, L., Zheng, W., Du, D., Zhang, Y., Ren, Y., Jiang, H., Cui, Z., Yu, H., Zhou, J., Zhang, S.: Authentic 4d driving simulation with a video generation model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28892–28902 (2025)
81. Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H.: Semantic image synthesis via diffusion models. arXiv preprint arXiv:2207.00050 (2022)
82. Wang, X., Zhu, Z., Zhang, Y., Huang, G., Ye, Y., Xu, W., Chen, Z., Wang, X.: Are we ready for vision-centric driving streaming perception? the asap benchmark. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9600–9610 (2023)
83. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., 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. pp. 14749–14759 (2024)
84. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6902–6912 (2024)
85. Wu, W., Guo, X., Tang, W., Huang, T., Wang, C., Chen, D., Ding, C.: Drivescape: Towards high-resolution controllable multi-view driving video generation. arXiv preprint arXiv:2409.05463 (2024)
86. Wu, Y., Xiang, Y., Tong, E., Ye, Y., Cui, Z., Tian, Y., Zhang, L., Liu, J., Han, Z., Niu, W.: Improving the robustness of pedestrian detection in autonomous driving with generative data augmentation. IEEE Network **38**(3), 63–69 (2024)
87. Xiao, A., Huang, J., Guan, D., Cui, K., Lu, S., Shao, L.: Polarmix: A general data augmentation technique for lidar point clouds. Advances in Neural Information Processing Systems **35**, 11035–11048 (2022)
88. Yan, T., Han, W., xia zhou, Zhang, X., Zhan, K., zhong Xu, C., Shen, J.: RLGF: Reinforcement learning with geometric feedback for autonomous driving video generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=EATkC9iHE3
89. Yan, T., Wu, D., Han, W., Jiang, J., Zhou, X., Zhan, K., Xu, C.z., Shen, J.: Drivingsphere: Building a high-fidelity 4d world for closed-loop simulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27531–27541 (2025)
90. Yang, K., Ma, E., Peng, J., Guo, Q., Lin, D., Yu, K.: Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. arXiv preprint arXiv:2308.01661 (2023)
91. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. Advances in Neural Information Processing Systems **36**, 18659–18675 (2023)
92. Yang, S., Yu, W., Zeng, J., Lv, J., Ren, K., Lu, C., Lin, D., Pang, J.: Novel demonstration generation with gaussian splatting enables robust one-shot manipulation. arXiv preprint arXiv:2504.13175 (2025)
93. Yang, X., Wen, L., Wei, T., Ma, Y., Mei, J., Li, X., Lei, W., Fu, D., Cai, P., Dou, M., et al.: Drivearena: A closed-loop generative simulation platform for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26933–26943 (2025)
94. Yang, Y., Liang, A., Mei, J., Ma, Y., Liu, Y., Lee, G.H.: X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. In: The Thirty-

ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=QclFsekj9B
95. Yang, Z., Yu, H., Feng, M., Sun, W., Lin, X., Sun, M., Mao, Z.H., Mian, A.: Small object augmentation of urban scenes for real-time semantic segmentation. IEEE Transactions on Image Processing **29**, 5175–5190 (2020)
96. Yang, Z., Guo, X., Ding, C., Wang, C., Wu, W., Zhang, Y.: Instadrive: Instance-aware driving world models for realistic and consistent video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25410–25420 (2025)
97. Yang, Z., Zhang, Y.: Consisdrive: Identity-preserving driving world models for video generation by instance mask. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=zgqFQM8VNe
98. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Yuxuan.Zhang, Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=LQzN6TRFg9
99. Yoon, Y., Cho, W., Ha, H., Kim, S., Yoon, K.J.: Expose: Reinforcing video generation models for extreme pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 32636–32646 (2026)
100. Yoon, Y., Jang, H.K., Yoon, K.J.: Gmt: Enhancing generalizable neural rendering via geometry-driven multi-reference texture transfer. In: European Conference on Computer Vision. pp. 274–292. Springer (2024)
101. Yoon, Y., Yoon, K.J.: Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12428–12438 (2023)
102. Yoon, Y., Yoon, K.J.: Resplat: Degradation-agnostic feed-forward gaussian splatting via self-guided residual diffusion. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=461VpgnLsi
103. Zeng, Y., Lin, Z., Zhang, J., Liu, Q., Collomosse, J., Kuen, J., Patel, V.M.: Scenecomposer: Any-level semantic image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22468–22478 (2023)
104. Zhang, J.O., Sax, A., Zamir, A., Guibas, L., Malik, J.: Side-tuning: a baseline for network adaptation via additive side networks. In: European conference on computer vision. pp. 698–714. Springer (2020)
105. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., et al.: Epona: Autoregressive diffusion world model for autonomous driving. arXiv preprint arXiv:2506.24113 (2025)
106. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
107. Zhang, X., Tseng, N., Syed, A., Bhasin, R., Jaipuria, N.: Simbar: Single image-based scene relighting for effective data augmentation for automated driving vision tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3718–3728 (2022)
108. Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., Huang, G., Chen, X., Wang, B., Zhang, Y., et al.: Drivedreamer4d: World models are effective data

machines for 4d driving scene representation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12015–12026 (2025)
109. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10412–10420 (2025)
110. Zhao, H., Ni, B., Fan, J., Wang, Y., Chen, Y., Meng, G., Zhang, Z.: Continual forgetting for pre-trained vision models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28631–28642 (2024)
111. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
112. Zhou, X., Liang, D., Tu, S., Chen, X., Ding, Y., Zhang, D., Tan, F., Zhao, H., Bai, X.: Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
113. Zhou, Y., Simon, M., Peng, Z., Mo, S., Zhu, H., Guo, M., Zhou, B.: Simgen: Simulator-conditioned driving scene generation. Advances in Neural Information Processing Systems **37**, 48838–48874 (2024)
114. Zhu, B., Wang, X., Li, H.: Consistentcity: Semantic flow-guided occupancy dit for temporally consistent driving scene synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26382–26392 (2025)
115. 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 Computer Vision and Pattern Recognition Conference. pp. 6812–6822 (2025)
116. Zhu, Z., Wu, Z., Zhu, Z., Zhou, L., Sun, H., WANG, B., Ma, K., Chen, G., Ye, H., Xie, J., jian Yang: Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=KWeX6tYno6