

ScenarioControl: Vision-language Controllable Vectorized Latent Scenario Generation

Lili Gao^{*1}, Yanbo Xu^{*2}, William Koch^{*2}, Samuele Ruffino¹, Luke Rowe³, Behdad Chalaki¹, Dmitriy Rivkin¹, Julian Ost^{1,2}, Roger Girgis^{1,3}, Mario Bijelic^{1,2}, and Felix Heide^{1,2}

¹ Torc Robotics, Blacksburg VA 24060, USA

² Princeton University, Princeton NJ 08544, USA

³ Mila, Montréal QC H2S 3H1, Canada

Abstract. We introduce ScenarioControl, the first vision-language control mechanism for learned driving scenario generation. Given a text prompt or an input image, ScenarioControl synthesizes diverse, realistic 3D scenario rollouts – including map, 3D boxes of reactive actors over time, pedestrians, driving infrastructure, and ego camera observations. The method generates scenes in a vectorized latent space that represents road structure and dynamic agents jointly. To connect multimodal control with sparse vectorized scene elements, we propose a cross-global control mechanism that integrates cross-attention with a lightweight global-context branch, enabling fine-grained control over road layout and traffic conditions while preserving realism. The method produces temporally consistent scenario rollouts from the perspectives different actors in the scene, supporting long-horizon continuation of driving scenarios. Extensive experiments validate that the control adherence and fidelity of ScenarioControl compare favorable to all tested methods across all experiments. To facilitate training and evaluation, we release a dataset with text annotations aligned to vectorized map structures at the project webpage: <https://light.princeton.edu/ScenarioControl>

1 Introduction

Large-scale datasets have been indispensable for the progress of autonomous driving [22], providing multi-modal labeled data across regions and conditions [2, 13, 23, 43, 44]. However, real-world logs alone are insufficient to capture rare but safety-critical events, such as wrong-way drivers. Evaluating these edge cases is essential for safe and reliable systems, yet relying solely on collected data is highly sample-inefficient [42].

Simulators bridge this gap by enabling safe, repeatable, and scalable experimentation. Traditional rule-based simulators such as CARLA [9], SMARTS [60], and MetaDrive [25] provide controllable virtual worlds for perception and planning research, yet their handcrafted world design limits realism and diversity,

* Equal contribution.

particularly for rare events [5, 27]. Recent generative approaches, such as Driving Diffusion [26] and Panacea [52], introduce controllability by conditioning on structured “control layouts” derived from existing scenarios which steer generation by projecting logged scenes into intermediate control representations [11, 21, 26, 29, 37, 41, 51, 52]. Behavior- and interaction-level methods control agent dynamics to create challenging closed-loop interactions, but commonly assume a given road graph and initial scene context [10, 15, 39, 62]. Even for a given initial scenario description, these methods often struggle to generate long-tail events as the datasets used to train them contain few of such examples. To the best of our knowledge, no existing method allows for vision-language-controlled scenario generation.

In parallel, diffusion-based and data-driven simulation work has made substantial progress on generating realistic and diverse driving scenes from data. These methods treat scenario synthesis as a generative modeling problem over vectorized [40] or rasterized representations [5, 7, 30, 36], either placing actors on an existing road graph [30, 36] or generating both the road graph and actor positions jointly [7, 40]. Generative models have been shown to produce high-fidelity structured outputs – road topology and traffic participants that are directly consumable by downstream components such as motion planning, sensor simulation, end-to-end driving, and multimodal future generation [14, 15, 28, 48, 53, 62]. However, a crucial gap remains: while realism and diversity are consistently improving, the controllability at the scenario-level of the layouts themselves is still limited. Generation is either entirely uncontrolled - by sampling from the latent space - or with control signals taken from existing scenarios, and extrapolation does not expose interpretable control knobs [7, 40]. We propose ScenarioControl to bridge this gap. ScenarioControl enables vision-language control of diffusion-based generation, unlocking controllable synthesis of structured driving scenarios and camera sensor simulation for the ego, and any other actor in the scene. Unlike other methods, it does so without relying on given logged scene configurations or auto-labeled control layouts. The resulting scenarios are plug-and-play with established simulators and autonomy stacks.

To this end, we introduce a novel cross-global control mechanism that conditions sparse vectorized scene tokens on dense features, from either a text prompt or a dashcam ego image, via two complementary branches: a cross-attention branch for fine-grained control and a lightweight global-context branch for capturing high-level scene intent. Conditioning on natural language and visual cues enables goal-directed scenario synthesis that both reflects real-world context and targets specific long-tail regimes. By steering road geometry, agent placement, and traffic conditions with a text prompt or a single ego dashcam-style image, generation moves beyond unconstrained sampling while preserving realistic structure and dynamics, crucial for synthesizing safety-critical cases and building targeted evaluation/training sets. In addition to enabling prompt- and image-conditioned scene generation, ScenarioControl also supports scene outpainting and long-horizon video continuation, maintaining temporal and visual consistency. We confirm ScenarioControl’s fidelity and controllability with quantita-

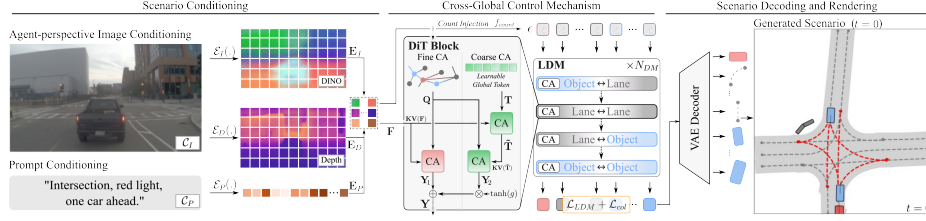


Figure 1: ScenarioControl Conditioning and Initial Scene Generation. Our method generates controllable, vectorized driving scenarios from visual (option 1) or prompt (option 2) conditioning. Text prompt or image embeddings guide the latent diffusion model (LDM) via cross-global control mechanism to ensure structurally valid scene generation (right). Red, blue, and grey tinting in both the latent and decoded spaces indicate ego vehicle, other vehicles, and lanes, respectively.

tive experiments, while qualitative results demonstrate adherence to conditioning and diverse long-horizon rollouts. Our contributions are summarized as follows:

- We propose a vision–language conditioned vectorized latent diffusion model that generates full vectorized driving scenes, including lane topology, agent placement, and traffic signals conditioned on text prompts or dash-cam-style ego images.
- We introduce a new conditioning mechanism that fuses sparse, vectorized road layouts with dense prompt and image representations, enabling fine-grained control over generation.
- We evaluate controllability, diversity, and fidelity of the generated scenarios for arbitrary actors in the scene, confirming that the proposed method performs favorably compared to existing methods while providing fine-grained vision-language control.

2 Related Work

The field of synthetic data generation for autonomous driving tasks can be clustered into *traffic simulation*, *scenario generation* methods and *sensor data generation* methods. The former focuses on behavioral simulation of traffic participants, whereas the latter encompasses the generation of maps, agent start locations, and static road obstacles. The last one builds on established behavior rollouts and scenario layouts to generate corresponding multi-modal sensor data.

Traffic Simulation. Traditional autonomous driving simulators such as CARLA [9] and others [3, 38] provide reproducible testbeds but rely on hand-crafted rules and scripted behaviors, limiting their ability to capture real-world driving diversity. Procedural generation approaches, including MetaDrive and SMARTS [25, 60], improve scalability but struggle with behavioral realism and rare event coverage. Data-driven simulators address these limitations by learning directly from real-world logs. Systems such as Waymax and GPUDrive [15, 24] enable hardware-accelerated training on logged scenarios, improving behavioral fidelity. However, replay-based approaches remain fundamentally constrained to observed patterns and cannot synthesize novel situations or explore counterfactual futures.

This limitation motivates generative simulators that synthesize new diverse and realistic scenarios, with several works addressing the behavior gap of log replay [10, 39]. TrafficGen [10] generates both initial agent states and agent behavior, CtRL-Sim [39] specializes in the latter. They use a transformer-based driving policy to enable controllable and adversarial agent behaviors but assume pre-existing scene layouts. Challenger [54] generates adversarial rollouts from fixed logged scenes for safety-critical evaluation, but does not generate editable vectorized scene structure.

Scenario Generation. Most traffic simulators require an initialized scene from which to roll out agent behaviors. A significant body of work focuses on producing suitable initial representations for agents [30, 32], lane graphs [33] or both [7, 40, 45], providing explicit structural control. When combined with traffic simulators, such initial scenes enable complete scenario generation with agent behaviors [7, 40]. SLEDGE [7] uses a raster-to-vector autoencoder with diffusion models to generate lane graphs and initial agent placements, but relies on rule-based traffic models for agent behaviors, limiting behavioral realism, while [40] builds on a vectorized scene representation to both accelerate the generation and increase quality. However, these methods do not allow for visual/textual grounding and focus on vectorized layouts for planners under flat-ground assumptions, rather than generating realistic sensor data.

Sensor Data Generation. A line of work focuses on reconstructing photorealistic sensor observations. Neural rendering methods based on NeRF and 3D Gaussian Splatting, including UniSim, NeuRAD, and Street Gaussians [46, 55, 57], achieve high visual fidelity but require explicit scene reconstruction and remain tethered to captured layouts, limiting their ability to generate counterfactual scenarios. Diffusion-based world models offer a more flexible alternative by directly synthesizing sensor videos. GAIA-1 [21] demonstrates controllable multi-camera generation, while subsequent work scales to longer horizons (LongDWM) and improves multi-view consistency (GAIA-2) [41, 50]. Cosmos-Drive-Dreams [37] extends these capabilities by post-training the Cosmos world foundation model on large-scale driving data, enabling precise HDMap and 3D bounding box control, multi-view expansion from single views, and LiDAR point cloud generation alongside RGB synthesis. DriveArena [56] explores closed-loop evaluation with diffusion models. Methods such as SimGen, GeoDrive and UMGGen [4, 53, 61] improve controllability by conditioning image generation on structural layouts and 3D geometry though these methods typically focus on rasterized map or non-structured scene representations. Methods that enable trajectory-level control, such as Epona [58] and ProphetDWM [49], or safety-critical scenario synthesis like AdvDiffuser [6], still operate either purely in abstract action control modes [12] or purely in pixel space.

ScenarioControl bridges this gap and exposes explicit control handles that allow for a controllable scenario generation with descriptive, interpretable scene prompts, subsequent behavioral roll-outs, and final sensor data generation. Unlike existing vectorized methods that generate scenes unconditionally, we enable the conditional generation of initial scenes. This produces complete vectorized

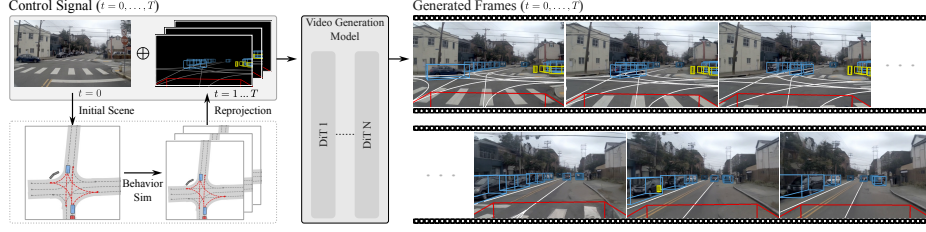


Figure 2: Scenario-Controlled Video Generation. Our method produces driving scene representations ($t = 0$) that we can simulate into scenario rollouts with a simulator, and use for downstream tasks such as video generation. The vectorized BEV scenario is projected into conditional camera-space layouts (BEV2Cam), and then fed into a LoRA-adapted Wan 2.2 (5B) model as a control signal for conditional video generation, together with the first frame.

scene graphs with explicit topology and agent placement that reflect real-world observations, or can support the generation of large-scale new datasets for specific scenarios using prompting. These structured representations can then drive traffic simulators for behavioral rollouts and serve as conditioning for photorealistic video generation, enabling full multi-modal scenario generation that maintains both structural consistency and visual realism. Where diffusion video models provide visual diversity without structured guarantees and vectorized simulators provide structural control without visual grounding, we offer both: controllable structured synthesis realized as photorealistic generation.

3 ScenarioControl

In the following, we first describe our scene representation and vectorized scene generation method (Sec. 3.1). Next, we describe the proposed conditioning mechanisms for camera observations or natural language text prompts (Sec. 3.2), also illustrated by Fig. 1. In Sec. 3.3 and Fig. 2, we describe scenario-guided video generation through vision-language control.

3.1 Scene Representation and Generation

We generate the initial scene by jointly modeling the underlying map structure and the actors’ initial states in a bird’s-eye-view (BEV) representation with elevation information. Specifically, we generate each scene within a $64\text{m} \times 64\text{m}$ field of view. A scene is then defined as a graph $\mathcal{I} = \{\mathcal{O}, \mathcal{M}\}$, comprising a set of objects $\mathcal{O}$ and a map structure $\mathcal{M} = \{\mathcal{L}, \mathbf{A}\}$. The map structure consists lane centerlines $\mathcal{L}$ and their connectivity graph $\mathbf{A}$. Our objective is to sample a scene $\mathcal{I} \sim p(\cdot | \mathcal{C})$ from some distribution p , conditioned on inputs $\mathcal{C}$, which can be either an image captured from the perspective of an agent in the scene or a text prompt describing the scene.

In contrast to existing scenario diffusion approaches that abstract the environment in a purely two-dimensional BEV representation [7, 40], we augment the object parametrization with vertical structure by incorporating elevation z and object height h . This provides essential elevation cues for downstream camera

and sensor-data synthesis, facilitating faithful reprojection of generated scenarios into the image domain as shown in Fig. 5.

3.2 ScenarioControl for Vectorized Latent Diffusion.

We propose a controllable vectorized latent diffusion model, as illustrated in Fig. 1, where we introduce the *cross-global control mechanism* that fuses sparse vector tokens with dense attention-based conditioning from prompts and images.

First, we train a transformer-based vectorized autoencoder that encodes scene elements into a compact latent representation. We then train a diffusion model with ϵ -predictor ϵ_θ over the latents of the autoencoder $\mathbf{Z} = [\mathbf{Z}_O, \mathbf{Z}_L]$ with variable cardinalities (N_o, N_l) . With condition $\mathcal{C}$, we minimize the DDPM ϵ -prediction objective [18] with

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\tau, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \mathcal{C}} \|\epsilon - \epsilon_\theta(\mathbf{Z}_\tau, \tau; \mathcal{C})\|_2^2, \quad (1)$$

where $\mathbf{Z}_\tau$ are noisy latents at step τ , composed of N_o object latents and N_l lane latents, and the noise vector decomposes as $\epsilon = (\epsilon_O, \epsilon_L)$. The conditioning inputs $\mathcal{C}$ comprise dense control signals $\mathcal{C}_I$ (*image*) and $\mathcal{C}_P$ (*prompt*) used in cross-attention, and the default control tokens $\bar{c}$ that encode the number of agents and lanes $N = (N_o, N_l)$ and the scene/domain label f (e.g., Singapore, Las Vegas, Boston, Pittsburgh in nuPlan), injected via AdaLN-Zero conditioning [35].

We next describe the proposed multi-modal control mechanism in more detail, which supports conditioning from both scene prompts and agent-perspective images.

Prompt Conditioning. Given a prompt describing the present scene and actors, we employ a text encoder $\mathcal{E}_P$ that extracts token embeddings as

$$\mathbf{E}_P = \mathcal{E}_P(\mathcal{C}_P) \in \mathbb{R}^{M_P \times d_P}, \quad (2)$$

where M_P is the number of prompt tokens and d_P is the embedding dimension of the text encoder. The embeddings $\mathbf{E}_P$ are then projected through an MLP layer with the *linear projection weights* $\mathbf{W}_P$ to obtain control tokens $\mathbf{F}_P \in \mathbb{R}^{M_P \times d}$, where d denotes the latent dimension used for agent or lane representations

$$\mathbf{F}_P = \mathbf{E}_P \mathbf{W}_P. \quad (3)$$

These prompt control tokens enter the control mechanism/attention by providing keys/values and are combined with the self-attention outputs.

Agent-Perspective Image Conditioning. Given a forward-facing agent-perspective image $\mathcal{C}_I$, e.g., a dashcam style image, we extract dense features with a vision backbone $\mathcal{E}_I$ and depth estimator $\mathcal{E}_D$ as

$$\begin{aligned} \mathbf{E}_{\text{feat}} &= \mathcal{E}_I(\mathcal{C}_I) \in \mathbb{R}^{M_I \times d_{\text{feat}}}, \\ \mathbf{E}_{\text{depth}} &= \mathcal{E}_D(\mathcal{C}_I) \in \mathbb{R}^{M_I \times d_{\text{depth}}}, \end{aligned} \quad (4)$$

where M_I denotes the number of image tokens, d_{feat} the feature dimension of the vision backbone, and d_{depth} the dimension of the estimated depth map. We

use pretrained models for both the vision backbone and the depth estimator, and both are frozen during training. Both image features and depth maps are projected to the model’s hidden dimension d via learned linear mappings. Further, we add a non-trainable sine-cosine positional embedding $\mathbf{P}$ to the image features [35] and depth maps, enforcing a shared spatial encoding across modalities

$$\mathbf{F}_m = \mathbf{E}_m \mathbf{W}_m + \mathbf{P}, \quad m \in \{\text{feat}, \text{depth}\}. \quad (5)$$

where $\mathbf{W}_{\text{feat}}$ and $\mathbf{W}_{\text{depth}}$ are learned *linear projection weights* mapping modality-specific features to the shared hidden dimension d , and $\mathbf{P}$ ensures alignment within the same image coordinate frame across both image features and depth maps. Finally, we concatenate these position-aware tokens to form the control feature representation $\mathbf{F}_I = [\mathbf{F}_{\text{feat}}, \mathbf{F}_{\text{depth}}]$.

Cross-Global Control Mechanism. Given the conditioning features, $\mathbf{F}_I$ and $\mathbf{F}_P$, we introduce a control mechanism to steer scenario generation. We employ N_{DM} factorized attention blocks, each applying (in order) object-to-lane, lane-to-lane, lane-to-object, and object-to-object self-attention (SA) over the stacked object and lane tokens. The conditioning inputs are injected into each (CA) component, which we detail in the following section. An overview of the full mechanism is illustrated in Fig. 1.

Conditioning vectorized scene tokens (lanes and agents) on agent-perspective images or scene prompt embeddings is inherently *unaligned*: a single scene token may depend on evidence from arbitrary subsets of conditioning tokens (e.g., occluded actors, distant lane cues, or globally specified textual constraints). Although cross-attention can, in principle, model such dependencies by allowing each query to attend to all keys, it provides little inductive structure and can be sample-inefficient when learning global context.

We therefore compute cross-attention through two parallel branches that share the key/value projections of the conditioning stream. Given scene queries $\mathbf{Q} \in \mathbb{R}^{N_q \times d}$ and conditioning tokens $\mathbf{F} \in \mathbb{R}^{N_k \times d}$, we first compute cross-attention

$$\mathbf{Y}_1 = \text{Attn}(\mathbf{Q}, \mathbf{K}(\mathbf{F}), \mathbf{V}(\mathbf{F})), \quad (6)$$

implemented efficiently with FlashAttention. In parallel, we introduce a small set of learned latent tokens $\mathbf{T} \in \mathbb{R}^{L \times d}$ that aggregate global context from $\mathbf{F}$, and expose this compact summary back to the scene queries

$$\mathbf{Y}_2 = \text{Attn}(\mathbf{Q}, \mathbf{K}(\tilde{\mathbf{T}}), \mathbf{V}(\tilde{\mathbf{T}})), \quad (7)$$

with

$$\tilde{\mathbf{T}} = \text{Attn}(\mathbf{T}, \mathbf{K}(\mathbf{F}), \mathbf{V}(\mathbf{F})), \quad (8)$$

with cost $\mathcal{O}(LN_k + N_q L)$. We combine both branches $\mathbf{Y} = \mathbf{Y}_1 + \tanh(g) \mathbf{Y}_2$ with a learned gate g . The gate is initialized such that $\tanh(g) \approx 0$ (i.e., $g = 0$), ensuring that the module initially reduces exactly to standard cross-attention. During

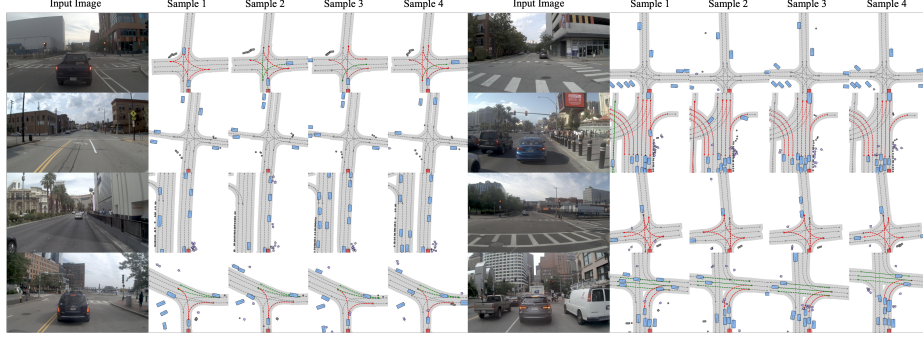


Figure 3: Dashcam-Style Image-Conditioned Scene Generation. Our model generates diverse initial scenes across different diffusion samples given a single input image. Elements with high certainty, such as vehicles in visible range, are found consistently across most samples. Elements with lower certainty, e.g., far away or outside the FOV, are sampled plausibly by our model.

training, the model can then progressively improve by selectively incorporating the additional global-context pathway. Since both pathways share the $\mathbf{K}(\cdot)$ and $\mathbf{V}(\cdot)$ projections, the parameter overhead remains minimal. Finally, the output is fused with multi-head self-attention via AdaLN-Zero modulation.

Count Injection f_{count} . Graph-based representations offer a natural handle for controlling scene complexity: the number of lanes and agents can be set directly by initializing the corresponding numbers of lane and object nodes. For instance, generating a two-lane highway can be guided by instantiating the matching number of lane nodes, while in the image/prompt-conditioned setting, these counts can also be inferred from the conditioning signal. We therefore train a lightweight attention-based regressor f_{count} that predicts the number of agents and lanes (N_o, N_l) from the conditioning tokens $\mathbf{F}$.

Compare to raster encodings or post-hoc control modules (e.g., ControlNet/T2I-Adapter [34, 59]), our method directly operates on variable-length vector tokens, thereby preserving topology (lane connectivity) and scene structure.

3.3 Video Generation and Continuation

The vectorized scene graph representation $\mathcal{I}$ enables direct BEV behavior simulation (with elevation) and camera-view video synthesis without an additional learned lifting step required by rasterized representations. It is used directly for behavior simulation and projected into the camera frame for video generation as control inputs to a video diffusion model, as described below.

Behavior Simulation. Real-world cameras capture far beyond the 64m initial layout. We use diffusion outpainting to construct large-scale scenes, which are temporally rolled out with a behavior model $\mathcal{BT}$ to generate diverse, consistent agent trajectories. This yields scenario rollouts $\mathcal{I}t = 1^T$, which are reprojected from BEV to camera space as wireframe sequences $\hat{\mathcal{C}}_{wire, t=1}^T$.

Sensor Video Generation. We adapt a video generation model $\mathcal{V}_i$ to generate photorealistic clips from the behavior simulation. We train two variants:

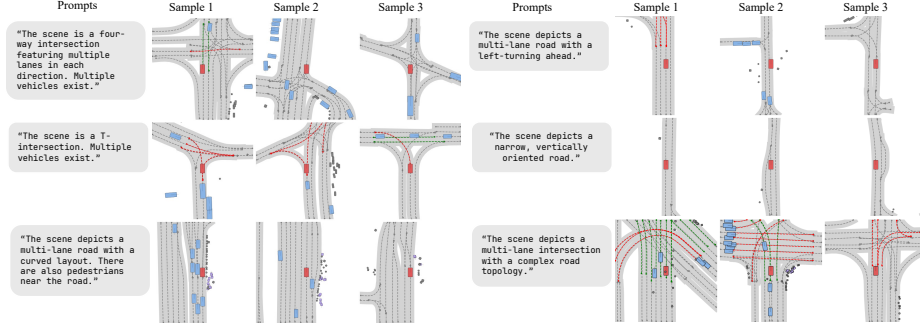


Figure 4: Prompt-Controlled Scenario Generation. Our model generates diverse initial scenes across different diffusion samples, with lane and object placement in each adhering to the text prompt. Text prompts are more open-ended than images, resulting in a larger variety across samples.

an image-conditioned model that uses the first frame C_I for appearance, and a prompt-conditioned model where the scene appearance is controlled by prompt description. Both are also trained to adhere to control sequences (wireframe renders of the behavior simulation) $\{\hat{C}_{\text{wire},t}\}_{t=1}^T$ to obtain photorealistic multi-frame renders $\{\hat{I}_t\}_{t=0}^T$ that follow the simulated agent behavior (see Fig. 2). Grounding video generation in the vectorized scene representation further enables novel-view synthesis by transforming camera extrinsics and re-rendering the wireframe representation from non-ego viewpoints, while keeping the text prompt fixed.

After adaptation, the model achieves fine-grained control of traffic behavior – either as part of a video continuation task with the first-frame conditioning or in completely novel traffic situations from a prompt. Examples of controlled rollouts are reported in Fig. 5 and 6, respectively.

Finally, as shown in Fig. 7, our approach enables the creation of challenging corner-case driving scenarios. Because the generated scenes are structured, they can be edited manually or programmatically while remaining compatible with downstream trajectory and video generation. We provide more details in the Supplementary Material (Supp.) Sec. 1.4 and the supplementary video.

3.4 Training

We adopt a two-stage training strategy. We first train the encoder f_e and decoder f_d to learn a reliable projection into the latent space. Next, we train the generative control mechanism and freeze the f_e, f_d weights and only train the dense condition blocks as described in Sec. 3.2, leveraging either the condition on Prompt or Image conditions.

Additionally, we also employ *classifier-free guidance* (CFG) [19]. Specifically, the conditioning inputs $\mathbf{F}_I$ and $\mathbf{F}_P$ are randomly dropped with probability p_{CFG} during training, encouraging the model to learn both conditional and unconditional behaviors. At inference, the guided prediction is computed as

$$\epsilon_{\text{CFG}} = \epsilon_{\theta}(\mathbf{Z}_{\tau}, \tau; \emptyset) + w[\epsilon_{\theta}(\mathbf{Z}_{\tau}, \tau; \mathcal{C}) - \epsilon_{\theta}(\mathbf{Z}_{\tau}, \tau; \emptyset)], \quad (9)$$

where w is the guidance weight controlling the strength of conditioning.

Scene Layouts. At test time, we support three generation modes reflecting the asymmetry between prompt and image conditioning. Text prompts can describe the full surroundings of the ego agent, whereas a single camera frame typically constrains only the forward-facing view. We define two 64×64 m canonical scene layouts: (1) an ego-centered crop $\mathcal{F}_P$, and (2) a forward-shifted crop $\mathcal{F}_I$, where the ego is placed near the edge to maximize forward coverage.

Crucially, we train the model on the same layout types used at inference, enabling consistent long-horizon rollouts. In practice, we use prompt-controlled synthesis for $\mathcal{F}_P$, image-conditioned completion for $\mathcal{F}_I$, and forward outpainting to extend either crop beyond the current field-of-view by sampling additional scene-graph nodes. Completion and outpainting are realized with masked denoising: latents corresponding to observed nodes are clamped, while the remaining tokens are sampled conditioned on $\mathcal{C}$.

Collision Penalty. In practice, conditioning on prompt and images can induce *cluttered* scene hypotheses: images contain occlusions and missing context, while prompt is often underspecified. Both effects can place multiple agents into the same plausible region, resulting in overlapping boxes in the decoded scene graph. To encourage physically consistent layouts, we add a collision penalty $\mathcal{L}_{\text{col}}$ during training. Concretely, we decode intermediate latents at selected timesteps and compute pairwise overlaps between predicted agents i, j . We define the collision regularization loss as

$$\mathcal{L}_{\text{col}} = \frac{1}{N} \sum_{i \neq j} \tanh\left(\frac{\text{overlap}(i, j)}{\zeta}\right), \quad (10)$$

where ζ controls smoothness. Since d is unreliable at low signal-to-noise ratios, we weight the penalty by $w_\tau = 1 - \sqrt{1 - \bar{\alpha}_\tau}$, emphasizing later diffusion steps (smaller τ) where the predicted geometry is more meaningful. This regularizer reduces agent overlap in the initial scene and improves global scene consistency under ambiguous conditioning. The differentiability is detailed in the Supp. Sec. 5.1.

Implementation Details are provided in the Supp., highlighting scene definitions, model architecture and training details .

4 Vision-language Scenario Dataset

To facilitate the training of our vision and language conditioned model, we curate a dataset consisting of images, natural language descriptions, and BEV maps. We select driving scenarios with corresponding camera captures from the nuPlan dataset [23]. To generate scene-level captions (e.g., “An intersection with a red light and multiple vehicles on the road. Pedestrians are standing on the sidewalk.”), we first render BEV visualization images for each scene and use a VLM (GPT-4.1-mini) to produce descriptive captions.

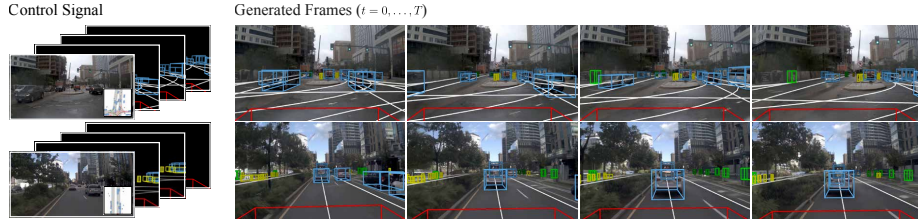


Figure 5: Controllable Video Continuations. Our video generation model respects control signals from the reprojected wireframe image sequences, and remains visually consistent with the initial frame. To show adherence with the control signal, we overlay the wireframe control signals over the generated sequences for qualitative evaluation. The samples shown are drawn from the NuPlan test split.

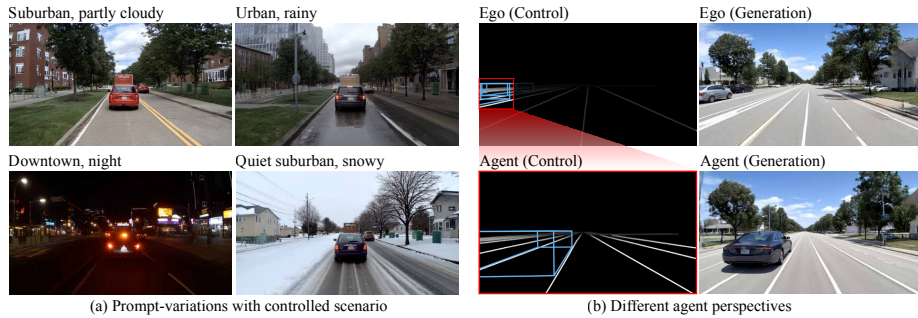


Figure 6: Prompt-conditioned Scenario Adaptations. The prompt-conditioned video generation variant allows (a) different scene variants given the same generated traffic rollout, and (b) generating the same traffic situation from different perspectives by transforming the camera pose while keeping the text prompt constant.

This process results in a large-scale dataset containing $\sim 500\text{K}$ high-quality captions, providing a rich foundation for training and evaluating our multimodal model. We provide additional details and dataset samples in the Supp.

5 Experiments

Next, we first introduce the relevant evaluation metrics in Sec. 5.1. We then compare our proposed Cross-Global Control conditioning mechanism with other popular conditioning mechanisms in Sec. 5.2. Subsequently, we demonstrate superior adherence to control input with significant overlap in scene content in Sec. 5.3. Next, we analyze the inclusion to predict the actor and lane counts, f_{count} , and to suppress collisions through the loss $\mathcal{L}_{col}$. Lastly, we provide a generalization experiment on the Waymo motion dataset [44] in Sec. 5.5. As our method focuses on *controllable* scenario generation, unconditioned comparison with prior methods [7, 10, 40] are orthogonal but can still be found in the Supp. Sec. 2.1.

5.1 Evaluation Metrics

Our evaluation focuses on two key aspects: generation quality and controllability. For generation realism, we adopt the same lane- and agent-level metrics as [40]

Table 1: Control Adherence and Realism Evaluation on Nuplan. We evaluate control adherence with three complementary metrics: cosine control similarity (CCS), shuffled perturbation gap (SPG), and control sensitivity correlation (CSC); and realism with Fréchet Distance (FD), Connectivity, Reach, and Convenience metrics. Across both settings, our method consistently improves control adherence and realism scores.

$\mathcal{F}_i$	Method	Global Control			Lane Control			Agent Control			Realism (Lane Topology)			
		CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$	CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$	CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$	FD $\downarrow$	Conn. $\downarrow$	Reach $\downarrow$	Conve. $\downarrow$
$\mathcal{I}$	Scenario Dreamer [40]	0.9896	0.0082	0.3510	0.4074	0.3568	0.2885	0.1909	0.1732	0.0625	2.44	0.22	0.12	0.73
	ScenarioControl	0.9908	0.0098	0.3809	0.4900	0.4367	0.4006	0.3163	0.3096	0.1626	0.51	0.32	0.11	0.35
$\mathcal{P}$	Scenario Dreamer [40]	0.9887	0.0141	0.4320	0.4400	0.3965	0.3166	0.1868	0.1800	0.0623	9.94	4.19	1.01	2.91
	ScenarioControl	0.9926	0.0194	0.6274	0.6606	0.6108	0.6100	0.3853	0.3865	0.2255	2.26	3.79	0.95	3.00

Table 2: Controllability Evaluation on the Waymo Motion dataset with prompt conditioning. The results show strong transfer across datasets.

$\mathcal{F}_i$	Method	Global Control			Lane Control			Agent Control		
		CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$	CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$	CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$
$\mathcal{I}$	Scenario Dreamer [40]	0.9922	0.0123	0.6701	0.3278	0.2775	0.2128	0.2264	0.2084	0.1220
	VLM+TrafficGen [10]	0.9564	0.0038	0.1969	0.0121	0.0352	0.0585	0.0181	0.0028	0.0054
	ScenarioControl	0.9950	0.0171	0.7587	0.6739	0.6204	0.5179	0.3586	0.3529	0.2112

Table 3: Analysis of Cross-Global Control Mechanism. We analyze the effect of our cross-global control mechanism with three complementary metrics: cosine control similarity (CCS), shuffled perturbation gap (SPG), and control sensitivity correlation (CSC). Across both image- and prompt-conditioned settings, our method consistently improves all three scores, indicating stronger adherence to the specified controls $\mathcal{F}_i$.

$\mathcal{F}_i$	Method	Global Control			Lane Control			Agent Control			Agent AP $\uparrow$	Collision RATE $\downarrow$
		CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$	CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$	CCS $\uparrow$	SPG $\uparrow$	CSC $\uparrow$		
$\mathcal{I}$	Concatenation	0.9898	0.0089	0.3606	0.4768	0.4286	0.3845	0.2786	0.2693	0.1340	23.03	24.43
	Full Cross-Attention [47]	0.9903	0.0093	0.3638	0.4803	0.4274	0.3830	0.2762	0.2722	0.1347	22.54	21.56
	Gated Attention [1]	0.9894	0.0082	0.3309	0.4013	0.3505	0.2695	0.1668	0.1624	0.0581	20.66	21.15
	Linear Attention [16]	0.9905	0.0086	0.3705	0.4416	0.3942	0.3505	0.2296	0.2303	0.0934	19.85	19.18
	AgentAttention [17]	0.9902	0.0091	0.3639	0.4827	0.4312	0.3893	0.2716	0.2667	0.1334	22.78	21.11
	SAAP Cross-Attention [31]	0.9904	0.0096	0.3628	0.4821	0.4278	0.3722	0.2556	0.2562	0.1264	22.03	27.51
	Windowed Attention [8]	0.9721	0.0078	0.1981	0.0899	0.1387	0.0893	0.0607	0.0655	0.0414	5.55	45.53
	Deformable Attention [63]	0.9802	0.0081	0.2503	0.3678	0.3301	0.2385	0.1439	0.1463	0.0525	21.65	23.62
	Squeezed Attention [20]	0.9882	0.0084	0.3404	0.4707	0.4202	0.3702	0.2021	0.2048	0.0869	23.97	20.00
	Cross-Global Control (Ours)	0.9908	0.0098	0.3809	0.4900	0.4367	0.4006	0.3163	0.3096	0.1626	23.36	20.07
$\mathcal{P}$	Concatenation	0.9881	0.0153	0.4105	0.5564	0.5116	0.4094	0.2501	0.2537	0.1314	23.37	30.18
	Full Cross-Attention [47]	0.9924	0.0191	0.6137	0.6596	0.6098	0.6075	0.3766	0.3788	0.2195	32.00	16.07
	Gated Attention [1]	0.9896	0.0156	0.4637	0.5795	0.5315	0.4953	0.2958	0.2945	0.1514	25.13	14.92
	Linear Attention [16]	0.9901	0.0177	0.5299	0.6390	0.5899	0.5861	0.3527	0.3500	0.2012	27.20	16.92
	AgentAttention [17]	0.9895	0.0159	0.4729	0.5963	0.5478	0.5110	0.3032	0.3051	0.1632	24.54	17.18
	SAAP Cross-Attention [31]	0.9907	0.0180	0.5414	0.5849	0.5336	0.4670	0.2531	0.2700	0.1318	26.05	18.17
	Windowed Attention [8]	0.9902	0.0180	0.5467	0.6356	0.5848	0.5644	0.2880	0.2905	0.1450	23.60	16.22
	Deformable Attention [63]	0.9883	0.0151	0.4546	0.4799	0.4415	0.3194	0.2058	0.2155	0.1016	21.60	13.82
	Squeezed Attention [20]	0.9901	0.0169	0.4927	0.6183	0.5675	0.5423	0.3167	0.3186	0.1689	28.25	19.94
	Cross-Global Control (Ours)	0.9926	0.0194	0.6274	0.6606	0.6108	0.6100	0.3853	0.3865	0.2255	33.46	17.07

and are presented in the Supp. For controllability we introduce a set of control metrics to assess adherence to the provided conditioning signals that build on top of Agent Accuracy, Collision Rate and Control Adherence.

Agent Accuracy. As a first step, we match generated agents to ground-truth agents and evaluate placement accuracy using average precision (AP). We compute AP_δ under center-point distance thresholds $\delta \in \{1.0, 2.0, 3.0\}$ m and report their mean, yielding an equally weighted measure of how well the control signal localizes agents. For the image-conditioned setting, AP is evaluated

only for agents within the camera field of view (FOV), whereas for the prompt-conditioned setting, AP is computed over all generated agents.

Collision Rate. We compute pairwise IoU over predicted agents and report the fraction of scenarios with intersecting bounding boxes.

Control Adherence. To quantify how well the predicted global, lane, and agent attributes follow the intended controls, we report three complementary metrics: Cosine Control Similarity (CCS), Shuffled Perturbation Gap (SPG), and Control Sensitivity Correlation (CSC), which are detailed in the Supp. **CCS** measures the alignment between the conditioning signal and the corresponding change in the generated scene, capturing direct controllability. **SPG** evaluates causal dependence by comparing the model’s response to correct versus randomly shuffled conditions; a larger gap indicates stronger conditional consistency. **CSC** measures the correlation between variations in the conditioning input and variations in the generated output, reflecting the sensitivity and smoothness of control. We evaluate across three levels (global, lane, and agent) to separately assess control over large-scale layout, map structure, and agents.

5.2 Analysis of Control Mechanism

We validate our cross-global control module by comparing it against a broad set of attention designs commonly used for fusing dense conditioning signals with token sequences. To our knowledge, this is the first study that systematically evaluates such mechanisms for conditioning a *vectorized 3D scene graph* on dense prompt or image features. Specifically, we compare against simple concatenation, full cross-attention [47], gated attention [1], linear attention [16], AgentAttention [17], SAAP cross-attention [31], windowed attention [8], deformable attention [63], and squeezed attention [20]. Quantitative results on lane- and agent-control metrics are reported in Tab. 3 with a qualitative result in Fig. 8.

Overall, attention-based conditioning mechanisms consistently outperform simple concatenation for both prompt- and image-conditioned control. In the image-conditioned setting, our method reduces collision rate by 43.4% and improves global CSC by 52.8% over concatenation; moreover, it also improves over full cross-attention with a 2.7% gain in agent CSC and a 4.6% gain in AP, while also achieving stronger global and lane control. In the prompt-conditioned setting, our method also surpasses all other baselines on most controllability metrics. This confirms that our cross-global attention captures the dense-to-sparse correspondences needed to align scene tokens with the a control input $\mathcal{F}_i$ better.

5.3 Analysis of Control Adherence

Fig. 3 and Fig. 4 compare controlled scene generation using image and prompt conditioning. Image inputs retain the visible scene structure and extend it with plausible unobserved regions, whereas text prompts define high-level attributes, such as intersection type, and therefore support broader sample diversity. Importantly, this added controllability does not come at the cost of realism: realism is preserved or improved for conditioned generation, as shown in Tab. 1 and Supp. Tab. 2. We additionally compare against the map-conditioned agent-placement approach in Supp. Secs. 2.2.

Table 4: Ablation of f_{count} and collision $\mathcal{L}$. The agent count can be specified a priori or inferred by f_{count} , showing correlation between both.

$\mathcal{L}_{col}$	f_{count}	Collision RATE ↓	Agent AP ↑
✗	✗	19.81	39.55
✓	✗	14.07	38.99
✗	✓	18.78	38.70
✓	✓	13.36	38.03



Figure 7: Corner cases. First row: a reversing white van. Second row: a traffic cone ahead of the ego vehicle, a child chasing a ball, and truck swerving into the lane.

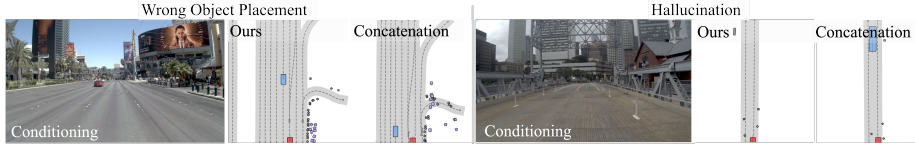


Figure 8: Qualitative comparison with Concatenation. Concatenation often violates the control input, e.g., missing object (left) or hallucinated agents (right).

Fig. 5 shows controllable video continuations that stay visually consistent with the initial view and follow the projected wireframe control signals over time. Finally, Tab. 1 reports higher CCS, SPG, and CSC across global, lane, and agent levels for the conditioned model, confirming improved control adherence and semantic alignment without compromising generation stability relative to the unconditioned model in [40].

5.4 Analysis of f_{count} and $\mathcal{L}_{col}$

In our vectorized representation, lane and agent counts can be specified when known or inferred by f_{count} ; Tab. 4 analyzes the impact of f_{count} and the collision loss on scene generation quality. The collision loss consistently reduces collisions in the generated scenarios, regardless of whether f_{count} is enabled. The estimated number of agents and lanes, however, strongly influences both Agent AP and the collision rate, reflecting a trade-off rooted in ambiguous conditioning: undercounting due to occlusions or limited FOV reduces Agent AP while also lowering the collision rate. Modeling a distribution over plausible cardinalities is a natural extension: sampling multiple counts per condition and selecting the most consistent one largely recovers Agent AP, leaving only a 1.5% gap from the oracle-count setting.

5.5 Generalization

We find that ScenarioControl generalizes across driving datasets to [44]. We evaluate transfer to the Waymo Open Motion dataset and report results in Tab. 2. By injecting explicit text prompts as control signals during generation, our model produces scenarios that better match the target distribution, outperforming Scenario Dreamer [40] across all reported metrics by up to 143%.

We additionally compare against a cascaded VLM+TrafficGen [10] baseline that first generates an intermediate BEV lane layout using GPT-4o mini and subsequently conditions TrafficGen on the generated map (Fig. 9). In contrast,

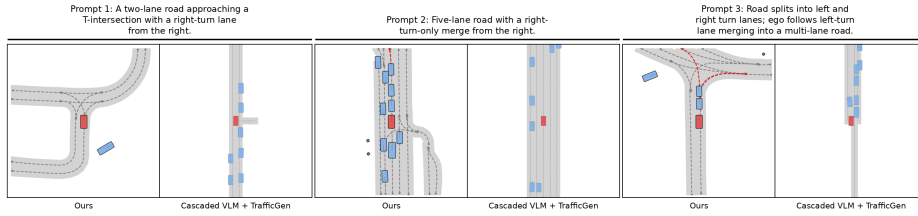


Figure 9: Qualitative comparison between our method and the cascaded VLM + TrafficGen baseline across three prompts.

ScenarioControl directly conditions lane and agent latents on vision-language features, avoiding error propagation through intermediate representations. The cascaded baseline increases latency to 13.05 s/sample due to costly VLM inference, while degrading generation quality, as shown in Fig. 9 and Tab. 2.

6 Conclusion

We introduce ScenarioControl, a multimodal conditional method for generating controllable and realistic driving scenarios with text prompts or images. We achieve prompt and visual conditioning through our proposed cross-global attention mechanism for vectorized scenarios that fuse dense multimodal cues with sparse map-agent structures. Additionally, the proposed method supports sensor video generation, producing temporally consistent renderings that visually ground the synthesized scenarios. Through extensive evaluations, we confirm that our control mechanism is favorable for prompt and image conditions, while allowing for diversity and fidelity, effectively bridging structured scenario simulation with realistic sensor-level inputs and human-interpretable guidance.

Our method opens several promising directions for further research. An interesting direction is to leverage controllable scene generation to help autonomous vehicles anticipate traffic situations beyond line of sight information by synthesizing plausible continuations of partially observed environments. Extending our method to full multi-view and long-horizon visual conditioning could further improve consistency. Finally, extending the latent representation to unify initial scene generation and traffic simulation could allow the capture of additional semantics such as intent, weather, or interaction cues – providing even finer control over scenario structure and agent behavior.

Acknowledgements

Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, a Amazon Science Research Award, and a Bosch Research Award. Felix Heide is a co-founder of Algolux (now Torc Robotics), Head of AI at Torc Robotics, and a co-founder of Cephia AI.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* **35**, 23716–23736 (2022)
2. 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)
3. Cai, P., Lee, Y., Luo, Y., Hsu, D.: Summit: A simulator for urban driving in massive mixed traffic. In: *2020 IEEE International Conference on robotics and automation (ICRA)*. pp. 4023–4029. IEEE (2020)
4. Chen, A., Zheng, W., Wang, Y., Zhang, X., Zhan, K., Jia, P., Keutzer, K., Zhang, S.: Geodrive: 3d geometry-informed driving world model with precise action control. *arXiv preprint arXiv:2505.22421* (2025)
5. Chen, D., Zhu, M., Yang, H., Wang, X., Wang, Y.: Data-driven traffic simulation: A comprehensive review. *IEEE Transactions on Intelligent Vehicles* **9**(4), 4730–4748 (2024). <https://doi.org/10.1109/TIV.2024.3367919>
6. Chen, X., Gao, X., Zhao, J., Ye, K., Xu, C.Z.: Advdiffuser: Natural adversarial example synthesis with diffusion models. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 4539–4549 (2023). <https://doi.org/10.1109/ICCV51070.2023.00421>
7. Chitta, K., Dauner, D., Geiger, A.: Sledge: Synthesizing driving environments with generative models and rule-based traffic. In: *European Conference on Computer Vision*. pp. 57–74. Springer (2024)
8. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691* (2023)
9. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: *Conference on robot learning*. pp. 1–16. PMLR (2017)
10. Feng, L., Li, Q., Peng, Z., Tan, S., Zhou, B.: Trafficgen: Learning to generate diverse and realistic traffic scenarios. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 3567–3575. IEEE (2023)
11. 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* (2025)
12. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems* **37**, 91560–91596 (2024)
13. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. *The international journal of robotics research* **32**(11), 1231–1237 (2013)
14. Guan, Y., Liao, H., Wang, C., Liu, X., Zhang, J., Li, Z.: World model-based end-to-end scene generation for accident anticipation in autonomous driving. *Communications Engineering* **4**(1), 144 (2025)
15. Gulino, C., Fu, J., Luo, W., Tucker, G., Bronstein, E., Lu, Y., Harb, J., Pan, X., Wang, Y., Chen, X., et al.: Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems* **36**, 7730–7742 (2023)

16. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: FLatten Transformer: Vision Transformer using Focused Linear Attention . In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5938–5948. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2023). <https://doi.org/10.1109/ICCV51070.2023.00548>, <https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00548>
17. Han, D., Ye, T., Han, Y., Xia, Z., Pan, S., Wan, P., Song, S., Huang, G.: Agent attention: On the integration of softmax and linear attention. In: Computer Vision – ECCV 2024: 18th European Conference. pp. 124–140. Springer-Verlag, Berlin, Heidelberg (2024)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models, 2020. URL <https://arxiv.org/abs/2006>
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance (2022), <https://arxiv.org/abs/2207.12598>
20. Hooper, C.R.C., Kim, S., Mohammadzadeh, H., Maheswaran, M., Zhao, S., Paik, J., Mahoney, M.W., Keutzer, K., Gholami, A.: Squeezed attention: Accelerating long context length LLM inference. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 32631–32652. Association for Computational Linguistics, Vienna, Austria (Jul 2025). <https://doi.org/10.18653/v1/2025.acl-long.1568>, <https://aclanthology.org/2025.acl-long.1568/>
21. 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 (2023), <https://arxiv.org/abs/2309.17080>
22. 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)
23. Karnchanachari, N., Geromichalos, D., Tan, K.S., Li, N., Eriksen, C., Yaghoubi, S., Mehdipour, N., Bernasconi, G., Fong, W.K., Guo, Y., et al.: Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 629–636. IEEE (2024)
24. Kazemkhani, S., Pandya, A., Cornelisse, D., Shacklett, B., Vinitzky, E.: Gpu-drive: Data-driven, multi-agent driving simulation at 1 million fps. arXiv preprint arXiv:2408.01584 (2024)
25. Li, Q., Peng, Z., Feng, L., Zhang, Q., Xue, Z., Zhou, B.: Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. IEEE transactions on pattern analysis and machine intelligence **45**(3), 3461–3475 (2022)
26. 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)
27. Li, Y., Yuan, W., Zhang, S., Yan, W., Shen, Q., Wang, C., Yang, M.: Choose your simulator wisely: A review on open-source simulators for autonomous driving. IEEE Transactions on Intelligent Vehicles **9**(5), 4861–4876 (2024)
28. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., Wang, X.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. CVPR (2024)

29. Lin, H., Guo, Z., Zhang, Y., Niu, S., Li, Y., Zhang, R., Cui, S., Li, Z.: Drivegen: Generalized and robust 3d detection in driving via controllable text-to-image diffusion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27497–27507 (2025)
30. Lu, J., Wong, K., Zhang, C., Suo, S., Urtasun, R.: Scenecontrol: Diffusion for controllable traffic scene generation. In: IEEE International Conference on Robotics and Automation (ICRA) (2024)
31. Mazaré, P.E., Szilvasy, G., Lomeli, M., Massa, F., Murray, N., Jégou, H., Douze, M.: Inference-time sparse attention with asymmetric indexing. arXiv preprint arXiv:2502.08246 (2025)
32. Meng, Y., Wu, H., Zhang, Y., Xie, W.: Scenegen: Single-image 3d scene generation in one feedforward pass. arXiv preprint arXiv:2508.15769 (2025)
33. Mi, L., Zhao, H., Nash, C., Jin, X., Gao, J., Sun, C., Schmid, C., Shavit, N., Chai, Y., Anguelov, D.: Hdmapgen: A hierarchical graph generative model of high definition maps. arXiv (2021)
34. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI’24/IAAI’24/EAAI’24, AAAI Press (2024). <https://doi.org/10.1609/aaai.v38i5.28226>, <https://doi.org/10.1609/aaai.v38i5.28226>
35. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
36. Pronovost, E., Ganesina, M.R., Hendy, N., Wang, Z., Morales, A., Wang, K., Roy, N.: Scenario diffusion: Controllable driving scenario generation with diffusion. Advances in Neural Information Processing Systems **36**, 68873–68894 (2023)
37. Ren, X., Lu, Y., Cao, T., Gao, R., Huang, S., Sabour, A., Shen, T., Pfaff, T., Wu, J.Z., Chen, R., et al.: Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. arXiv preprint arXiv:2506.09042 (2025)
38. Rong, G., Shin, B.H., Tabatabaee, H., Lu, Q., Lemke, S., Mozeiko, M., Boise, E., Uhm, G., Gerow, M., Mehta, S., et al.: Lgsvl simulator: A high fidelity simulator for autonomous driving. In: 2020 IEEE 23rd International conference on intelligent transportation systems (ITSC). pp. 1–6. IEEE (2020)
39. Rowe, L., Girgis, R., Gosselin, A., Carrez, B., Golemo, F., Heide, F., Paull, L., Pal, C.: Ctrl-sim: Reactive and controllable driving agents with offline reinforcement learning. arXiv preprint arXiv:2403.19918 (2024)
40. Rowe, L., Girgis, R., Gosselin, A., Paull, L., Pal, C., Heide, F.: Scenario dreamer: Vectorized latent diffusion for generating driving simulation environments. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17207–17218 (2025)
41. 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)
42. Scanlon, J.M., Kusano, K.D., Daniel, T., Alderson, C.J., Ogle, A., Victor, T.: Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. Accident Analysis and Prevention **163**, 106454 (2021), <https://api.semanticscholar.org/CorpusID:232285642>
43. Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H.,

- Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
44. Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
 45. Sun, S., Gu, Z., Sun, T., Sun, J., Yuan, C., Han, Y., Li, D., Ang, M.H.: Drivescenes: Generating diverse and realistic driving scenarios from scratch. *IEEE Robotics and Automation Letters* **9**(8), 7007–7014 (2024). <https://doi.org/10.1109/LRA.2024.3416792>
 46. Tonderski, A., Lindström, C., Hess, G., Ljungbergh, W., Svensson, L., Petersson, C.: Neurad: Neural rendering for autonomous driving (2024)
 47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
 48. Wang, T., Zhang, C., Qu, X., Li, K., Liu, W., Huang, C.: Diffad: A unified diffusion modeling approach for autonomous driving. *ArXiv abs/2503.12170* (2025), <https://api.semanticscholar.org/CorpusID:277066594>
 49. Wang, X., Peng, P.: Prophetdwm: A driving world model for rolling out future actions and videos. *arXiv preprint arXiv:2505.18650* (2025)
 50. Wang, X., Wu, Z., Peng, P.: Longdwm: Cross-granularity distillation for building a long-term driving world model. *arXiv preprint arXiv:2506.01546* (2025)
 51. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: *European conference on computer vision*. pp. 55–72. Springer (2024)
 52. 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)
 53. Wu, Y., Zhang, H., Lin, T., Huang, L., Luo, S., Wu, R., Qiu, C., Ke, W., Zhang, T.: Generating Multimodal Driving Scenes via Next-Scene Prediction . In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 6844–6853. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2025). <https://doi.org/10.1109/CVPR52734.2025.00642>, <https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00642>
 54. Xu, Z., Li, B., Gao, H.a., Gao, M., Chen, Y., Liu, M., Yan, C., Zhao, H., Feng, S., Zhao, H.: Challenger: Affordable adversarial driving video generation. *arXiv preprint arXiv:2505.15880* (2025)
 55. Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., Peng, S.: Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In: *European Conference on Computer Vision*. pp. 156–173. Springer (2024)
 56. Yang, X., Wen, L., Ma, Y., Mei, J., Li, X., Wei, T., Lei, W., Fu, D., Cai, P., Dou, M., Shi, B., He, L., Liu, Y., Qiao, Y.: 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)

57. Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.C., Yang, A.J., Urtasun, R.: Unisim: A neural closed-loop sensor simulator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1389–1399 (2023)
58. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., Cao, X., Yin, W.: Epona: Autoregressive diffusion world model for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
59. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
60. Zhou, M., Luo, J., Villella, J., Yang, Y., Rusu, D., Miao, J., Zhang, W., Alban, M., Fadakar, I., Chen, Z., Huang, A.C., Wen, Y., Hassanzadeh, K., Graves, D., Chen, D., Zhu, Z., Nguyen, N., Elsayed, M., Shao, K., Ahilan, S., Zhang, B., Wu, J., Fu, Z., Rezaee, K., Yadmellat, P., Rohani, M., Nieves, N.P., Ni, Y., Banijamali, S., Rivers, A.C., Tian, Z., Palenicek, D., bou Ammar, H., Zhang, H., Liu, W., Hao, J., Wang, J.: Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving (11 2020), <https://arxiv.org/abs/2010.09776>
61. 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)
62. Zhou, Z., HU, H., Chen, X., Wang, J., Guan, N., Wu, K., Li, Y.H., Huang, Y.K., Xue, C.J.: BehaviorGPT: Smart agent simulation for autonomous driving with next-patch prediction. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), <https://openreview.net/forum?id=GRmQjLzaPM>
63. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (ICLR) (2021)