Challenger: Affordable Adversarial Driving Video Generation for Safety Testing

Zhiyuan Xu[1,2,*], **Bohan Li**[3,4,*], **Huan-ang Gao**[1], **Mingju Gao**[1]
Xin Jin[4], **Hang Zhao**[5], **Shuo Feng**[6], **Ya-Qin Zhang**[1], **Hao Zhao**[1,7,†]
[1]AIR, Tsinghua, [2]UCAS, [3]SJTU, [4]EIT, Ningbo
[5]IIIS, Tsinghua, [6]Department of Automation, Tsinghua, [7]BAAI
[*]Equal contribution. [†]Corresponding author.

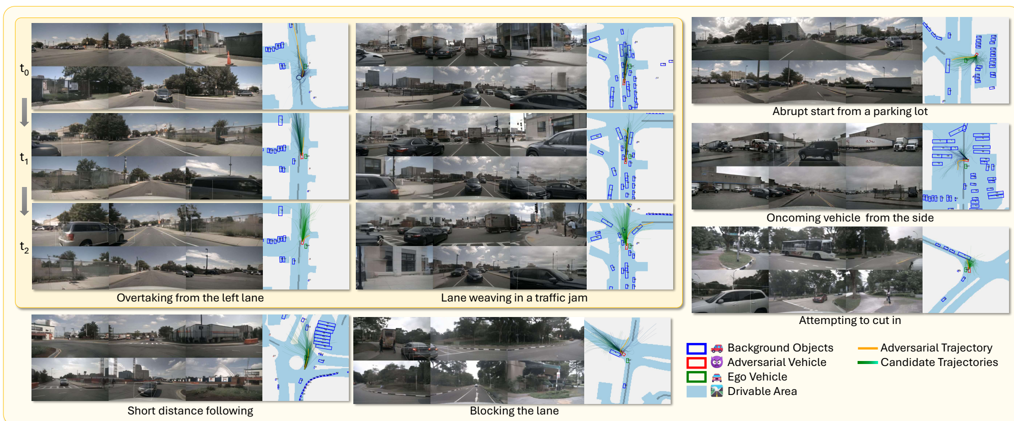


Figure 1: **Photorealistic adversarial driving videos generated by *Challenger*.** Each scenario includes an adversarial vehicle—highlighted with a white 3D bounding box in camera views—that is designed to challenge the ego vehicle through aggressive or unexpected maneuvers. *Challenger* autonomously produces these videos with an affordable computation budget. Readers are suggested to zoom in for detailed inspection. Please refer to our project page for video demonstrations.

Abstract

Ensuring the safety of autonomous driving (AD) systems requires rigorous evaluation under rare but challenging interactions. Recent photorealistic driving video generators largely model everyday traffic, while many adversarial scenario generation methods operate only at trajectory/BEV levels and do not yield sensor data that can stress-test end-to-end (E2E) AD models. In this work, we introduce **Challenger**[1], a framework that generates physically plausible yet photorealistic adversarial driving videos to support safety testing, and the collection of evidence for safety arguments. The core difficulty is to jointly optimize over the space of traffic interactions and high-fidelity sensor observations. Challenger makes this affordable through: (1) a physics-aware multi-round trajectory refinement process that narrows down candidate adversarial maneuvers, and (2) a tailored trajectory scoring function that encourages realistic yet adversarial behavior while maintaining compatibility with downstream video synthesis. As tested on the nuScenes

[1]The name *Challenger* refers to the framework's role in challenging autonomous driving models with difficult scenarios to probe robustness and expose vulnerabilities. It does not reference any specific vehicle or entity.

dataset, Challenger produces diverse adversarial behaviors (e.g., cut-ins, sudden lane changes, tailgating, blind-spot intrusions) and renders them as photorealistic multiview videos. Extensive evaluations show that these scenarios significantly increase the collision rate of state-of-the-art E2E AD models, with failures often transferring across models. By providing controlled, repeatable, photorealistic adversarial exposures, Challenger offers a practical instrument for safety testing, pre-deployment audits, and continuous safety monitoring. Our code, model, and dataset can be found at `https://pixtella.github.io/Challenger/`

1 Introduction

Autonomous driving systems operate in safety-critical environments where failures can cause harm and erode public trust [37, 25]. Ensuring that autonomous driving (AD) systems are safe and trustworthy requires extensive testing beyond average-case performance [34, 40]. However, these systems are typically trained and evaluated on large-scale datasets such as nuScenes [7], OpenScene [12], NAVSIM [13] and Bench2Drive [26], which predominantly reflect natural traffic flows and lack the deliberately challenging interactions needed to probe safety margins and expose vulnerabilities.

To address this gap, researchers have explored adversarial or risk-targeted scenario generation. For example, Feng et al. [17] model adversarial behavior in closed-loop simulation to analyze risk patterns under controlled conditions; Ding et al. [15] synthesize rare-event distributions but limit interactions to two vehicles and do not produce corresponding sensor data; NeuroNCAP [35] offers closed-loop photorealistic testing but depends on manually constructed scenarios confined to a few predefined adversarial categories. Photorealistic driving video generation has also advanced but still focuses on everyday behaviors, leaving safety-relevant edge cases underexplored. In short, a key capability remains missing: automatically generating diverse, physically plausible, photorealistic adversarial driving videos in real-world scenes for systematic AD safety testing and auditing.

To fill this gap, we introduce *Challenger*, a framework that synthesizes high-fidelity multiview videos where an adversarial agent challenges the ego vehicle in realistic urban scenes. Figure 1 shows generated adversarial scenarios. Each includes an ego vehicle (green) and an adversarial vehicle (red), embedded within realistic traffic scenes replayed from nuScenes. The adversarial agent executes physically plausible yet aggressive maneuvers—such as sudden cut-ins, blind-spot intrusions, tailgating, and lane blocking—that are designed to interfere with the ego vehicle's driving decisions. While challenging, these scenarios remain solvable, ensuring meaningful safety tests of AD models. They cover diverse urban conditions, including multi-lane traffic, intersections, roundabouts, and parking areas. Beyond handcrafted corner cases, they provide an affordable automated approach for assessing how autonomous driving models respond to unexpected and potentially unsafe behaviors. Importantly, this is not a substitute for real-world testing, and the generated videos are not intended for deployment training.

In summary, our key contributions are as follows:

- We introduce the task of adversarial driving video generation and propose *Challenger*, a unified framework that makes it affordable via multi-round physics-aware trajectory refinement and renderer-compatible adversarial scoring.
- We validate *Challenger* on the nuScenes dataset, generating diverse and photorealistic adversarial maneuvers—such as cut-ins, tailgating, and blocking—across a wide range of dynamic traffic contexts including intersections, roundabouts, and parking lots.
- Videos generated by *Challenger* significantly degrade the performance of state-of-the-art AD models (UniAD [23], VAD [27], SparseDrive [45], DiffusionDrive [33]) and reveal transferable failure patterns, supporting risk analysis and incident-prevention insights.

2 Related Work

2.1 Generating Driving Trajectories

Driving trajectory generation, prediction, and adversarial manipulation are central to AD safety validation. Recent techniques [17, 16, 55, 57, 63, 46, 52, 32, 39] enhance scenario diversity, realism, and controllability. Feng et al. [17] integrate naturalistic and adversarial behaviors to reduce evaluation mileage without sacrificing fairness. D2RL [16] densifies safety-critical events from real data to accelerate validation. To increase realism, NeuralNDE [55] models multi-agent interactions and rare-event conflicts via learned critics and safety mapping networks. Diffusion-based planners have also emerged as effective trajectory generators. Diffusion-ES [57] employs gradient-free optimization for instruction-conditioned synthesis, while Zheng et al. [63] introduce a transformer variant for multi-modal behavior and generalization across driving styles. In parallel, language-conditioned methods enable intuitive scenario customization. LCTGen [46], InteractTraj [52], and ChatSUMO [32] employ LLMs for traffic scene and trajectory generation with controllable semantics. VLP [39] integrates language and planning for robust long-tail scene understanding. These works typically lack photorealistic sensor data or explicit adversarial intent. Challenger bridges this gap by combining diffusion-based trajectory generation, physics-aware planning, and multiview rendering to produce photorealistic adversarial driving videos that systematically stress-test AD models.

2.2 Driving Video Generation

Driving video generation synthesizes camera sensor observations, which have become a primary sensing modality for many contemporary AD systems [1]. Recent advancements in this domain have enhanced the realism and controllability of autonomous driving simulations [20, 21, 29, 19, 31, 51, 36, 48]. DriveDreamer [50] improves video and action prediction accuracy by enforcing real-world traffic constraints. Its successor, DriveDreamer-2 [61], integrates large language models to enable user-defined video generation with enhanced temporal coherence. MagicDrive [20] achieves high-fidelity street-view synthesis via tailored encoding and cross-view attention mechanisms. Extensions like MagicDriveDiT [19] and MagicDrive3D [18] address scalability through DiT architectures and deformable Gaussian splatting with monocular depth initialization, respectively. UniScene [30] unifies multi-modal data generation (e.g., occupancy, RGB, LiDAR) using progressive processes, while DreamDrive [36] synthesizes 4D scenes via hybrid Gaussian representations for balanced visual quality and generalizability. Drive-WM [51] pioneers multiview video forecasting by decoupling spatial-temporal dynamics with view factorization. Stag-1 [48] advances 4D simulation through spatial-temporal decoupling and point cloud reconstruction, achieving photo-realistic, viewpoint-agnostic scene evolution. Meanwhile, StreetCrafter [56] and FreeVS [49] leverage diffusion models for controllable novel-view synthesis, expanding capabilities for urban scene generation. These methods primarily simulate natural traffic flows. In contrast, Challenger specifically targets adversarial videos, combining diffusion-based trajectory optimization with multiview rendering to expose model vulnerabilities under physically plausible yet deliberately challenging conditions.

2.3 End-to-End Autonomous Driving

End-to-end autonomous driving (E2E-AD) [5, 38, 62, 10, 23, 54, 58, 53] integrates perception, decision-making, and control into a unified, holistically optimized framework. Pioneering work by Bojarski et al. [5] demonstrated direct mapping from raw visual inputs to steering commands using convolutional neural networks, bypassing modular pipelines. Subsequent research enhanced robustness through uncertainty quantification [38] and adversarial training [9], with the latter introducing a two-stage paradigm where privileged agents guide vision-based policy learning, achieving state-of-the-art results on CARLA/NoCrash benchmarks.

Recent advances emphasize multimodal trajectory prediction and interaction modeling. Methods like R-Pred [11], EigenTrajectory [3], and SingularTrajectory [4] exemplify this trend. Safety validation techniques, including adversarial 3D object optimization [43] and traffic condition manipulation [60], generate challenging scenarios via closed-loop simulations. Neural rendering [2] further extends adversarial testing by perturbing scene textures. Scalability and uncertainty management are addressed through coverage-driven testing [47] and probabilistic trajectory prediction [28]. Architectural innovations, such as SparseAD [59] and GenAD [62], improve multi-task efficiency. Attention mechanisms [10] and multimodal fusion frameworks [42, 58] enhance interpretability and vehicle-to-everything (V2X) coordination, with UniAD [23] achieving top performance via task-aware planning.

Despite progress, E2E-AD systems remain vulnerable to adversarial attacks [6]. Mitigations include reinforcement learning for sensor failure resilience [24] and large language model integration for interpretable control [54]. OpenEMMA [53] promotes accessibility through multimodal toolkits.

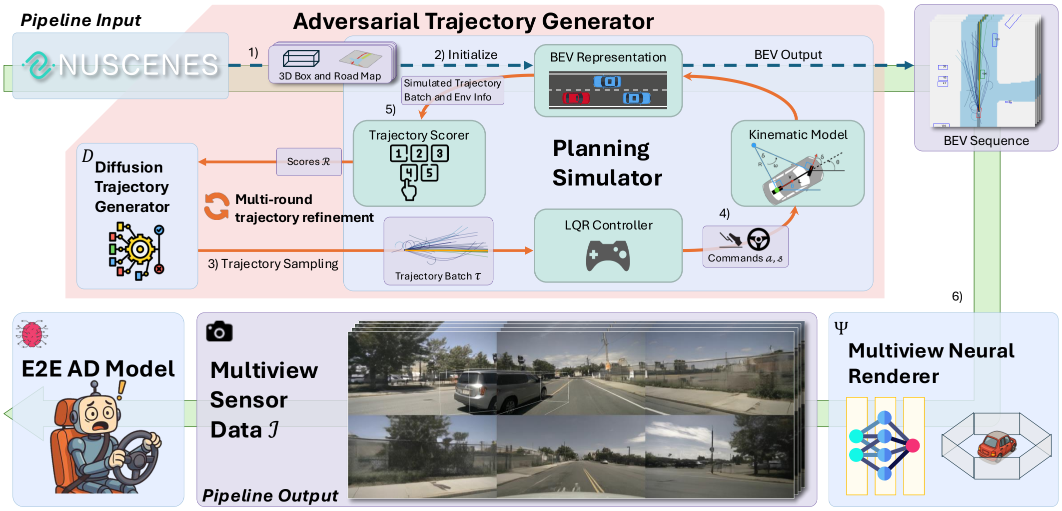


Figure 2: ***Challenger* System Overview.** 1) *Challenger* first takes 3D bounding boxes and BEV road maps from a real-world dataset (e.g., nuScenes) to initialize a driving scene. 2) It then randomly selects a background vehicle to act as the adversarial agent, while leaving all other participants unchanged. 3) At fixed keyframes, *Challenger* plans the adversarial vehicle's trajectory and executes it continuously between keyframes. At each planning keyframe, a multi-round refinement process is used to explore the trajectory space efficiently and generate an adversarial one. 4) To ensure physical feasibility, a physics-aware planning simulator with an LQR controller and a kinematic model is used to simulate the motion of the adversarial vehicle. 5) The simulated trajectories are then scored, and the top-performing trajectories are resampled (with replacement), perturbed by noise, denoised via the diffusion model, and fed back into the planning simulator for subsequent rounds. 6) Finally, a multiview neural renderer synthesizes photorealistic video outputs of the final adversarial scenario.

These models are evaluated under benign conditions and remain vulnerable to rare or adversarial events. Challenger addresses this limitation by synthesizing photorealistic adversarial videos that expose E2E-AD failure cases, offering a practical, scalable complement to existing evaluation pipelines.

3 Method

Our proposed *Challenger* comprises three key components: a diffusion-based trajectory generator (Section 3.1) that enables multi-round trajectory refinement (Section 3.4), a physics-aware planning simulator (Section 3.2) with a trajectory scorer (Section 3.3), and a multiview neural renderer (Section 3.5). Figure 2 provides an overview.

3.1 Diffusion Trajectory Generator

Diffusion Model Preliminaries. Diffusion models [14, 22, 44] are a class of generative models that learn to synthesize data by reversing a gradual noising process. During training, a forward diffusion process incrementally adds Gaussian noise to clean data over T steps, transforming an initial sample $\tau_0 \sim p(\tau_0)$ into a noise distribution $\tau_T \sim \mathcal{N}(0, I)$. The model is trained to approximate the reverse process by learning a denoising function $p_\theta(\tau_{t-1}|\tau_t)$, which estimates the distribution of the previous step given the current noisy input. To generate new data, the sampling process begins with pure noise τ_T and iteratively applies the learned denoising function over T steps, ultimately producing a sample τ_0 from the target data distribution. **In the case of trajectory generation**, diffusion models are promising because they can learn the distribution of driving trajectories in a data-driven manner, and are capable of generating diverse and complex driving behaviors.

Naturalness before Adversary. Inside *Challenger*, before the adversarial vehicle can learn to drive adversarially, it must first acquire the fundamental ability to generate realistic driving behavior. To this end, we train an *unconditional* diffusion model that captures the distribution $p(\tau)$ of naturalistic driving trajectories, where each trajectory $\tau \in R^{T\times 2}$ is represented as a fixed-length sequence of 2D waypoints in the ego vehicle's local coordinate system. The model is trained on the nuPlan dataset [8], which provides a larger and more diverse set of real-world trajectories than nuScenes. This enables the adversarial agent to implicitly learn the underlying distribution of human-like motion patterns. Once trained, the diffusion model can generate plausible driving trajectories from noise via iterative denoising. It can also refine existing trajectories through a truncated reverse process—adding partial noise and re-denoising—yielding trajectory variants with controlled stochasticity. Placed inside the Adversarial Trajectory Planner of Figure 2, its ability forms the foundation of the adversarial vehicle's learned behavior, enabling it to explore and execute diverse yet physically plausible maneuvers in downstream planning.

3.2 Physics-aware Planning Simulator

While a trained diffusion model enables *Challenger* to sample diverse driving plans, they are not guaranteed to be physically feasible for the selected background vehicle, nor inherently adversarial to the ego vehicle. This is due to: (1) the selected background vehicle may have different size and dynamics compared to the vehicles in the nuPlan dataset, where the model is trained; and (2) the generated trajectories are sampled without explicit awareness of the surrounding driving context.

Guaranteeing physical feasibility. Therefore, we introduce a physics-aware planning simulator inside the Adversarial Trajectory Planner of Figure 2. It first uses a linear-quadratic regulator (LQR) controller to track each sampled trajectory by generating control commands (steering angle and acceleration) for the adversarial vehicle, and then simulates the motion of the adversarial vehicle using a kinematic bicycle model [41], which captures essential vehicle dynamics and produces the resulting 3D bounding boxes over time. Parameters of the model are customized to match the selected adversarial vehicle's characteristics.

3.3 Trajectory Scorer

Gathering context. In order to incorporate contextual information and find the most adversarial trajectory, we design a scoring function. Ideally, it needs to distinguish contextually feasible trajectories from those that are not, while also identifying the most adversarial ones. Although it is theoretically possible to render the resulting driving scenes for every trajectory in the batch and use a surrogate AD model to find which trajectory is the most adversarial (resulting in the highest failure rate of the AD model), this approach is computationally expensive and impractical. **Instead**, we propose a scoring function that operates at an abstract level and thus can be computed efficiently. Specifically, a composite score is calculated based on three criteria: (1) drivable area compliance, which penalizes deviation from the legal road surface; (2) collision rate, which penalizes contact with static or dynamic objects, including the ego vehicle; and (3) adversarial challenge, which rewards close encounters with the ego vehicle that increase the likelihood of failure without directly causing a collision. We place it right after physical simulation modules in the Adversarial Trajectory Planner of Figure 2 to provide feedback for trajectory selection or refinement in Section 3.4.

3.4 Multi-round Trajectory Refinement

Why being affordable matters. Finding high-quality adversarial trajectories requires searching the trajectory space, but naively sampling a huge batch and picking the best one is computationally expensive. Instead, *Challenger* employs a multi-round refinement process inspired by diffusion-based planning [57]. This iterative approach balances exploration and exploitation, enabling the adversarial vehicle to find strategically challenging trajectories affordably, without a prohibitively large batch size.

In the initial round, a batch of candidate trajectories $\tau \in R^{B\times T\times 2}$ is sampled from the diffusion model, simulated using the physics-aware planning simulator, and then scored. Subsequently, a resampling step selects trajectories from the current batch with replacement, assigning higher probabilities to those with superior scores. This results in a new batch $\tilde{\tau} \in R^{B\times T\times 2}$ that emphasizes promising candidates. To introduce diversity and facilitate exploration, Gaussian noise is added to these trajectories, and the diffusion model denoises them, producing a mutated batch $\hat{\tau}$ for the next round.

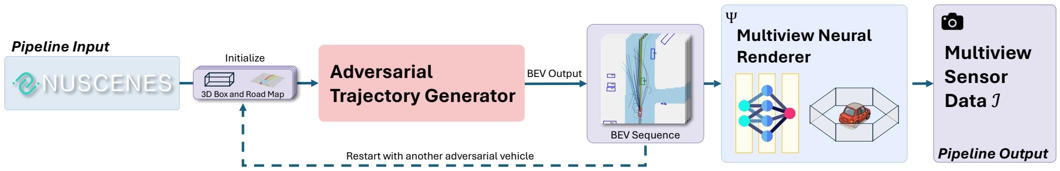


Figure 3: **Extending to Multiple Adversarial Vehicles.** Challenger can be repeatedly applied to the same initialized scene, sequentially designating different background vehicles as adversaries.

Challenging but solvable. This iterative process is conducted over a predetermined number of rounds. In each iteration, the refinement progressively generates trajectories that are increasingly tailored to challenge the ego vehicle while adhering to environmental constraints. Please refer to Figure 7 for an intuitive illustration. The trajectory yielding the highest final score is selected as the optimal adversarial plan. Importantly, in the open-loop evaluation setting, adversarial scenarios are designed to induce test-time failures of the AD model under test, rather than encoding inevitable collisions into the scenario itself. To ensure this distinction, *Challenger* verifies that the final trajectory avoids collisions with any objects and that the ego vehicle can safely navigate the scene by following its original path in the dataset. If these conditions are not met, *Challenger* **discards** the scene entirely. This filtering guarantees that every generated scenario presents a **challenging yet solvable** situation for the ego vehicle.

System integration before rendering. Together, the diffusion-based trajectory generator, physics-aware planning simulator, and multi-round refinement process enable *Challenger* to produce diverse, complex adversarial driving scenarios efficiently and affordably: (1) the diffusion model captures a wide spectrum of driving behaviors, generating varied trajectories that challenge the ego vehicle in multiple ways; (2) adversarial trajectory planning operates at an abstract level without costly video rendering; and (3) multi-round refinement progressively narrows and enhances candidates, creating high-quality adversarial scenarios without human supervision.

3.5 Multiview Neural Renderer

To synthesize photorealistic, multiview driving videos for our generated scenarios, we need a neural renderer after the Adversarial Trajectory Generator of Figure 2. The renderer should take BEV maps and 3D bounding boxes of objects as conditional input, and generate photorealistic videos from six cameras surrounding the ego vehicle. We use MagicDriveDiT [19], a diffusion-based model trained on nuScenes that can generate high-fidelity videos long enough to cover the full driving scene. Quantitative and qualitative results in Table 1, Table 3, and Figure 9 validate this choice.

3.6 Extending to Multiple Adversarial Vehicles

While our primary focus is single-adversary scenarios, *Challenger* naturally extends to multiple adversarial vehicles in the same scene (Figure 3). Before rendering, we rerun *Challenger* on the generated scene and designate another vehicle as adversarial, treating previously planned vehicles as part of the environment; this can repeat as needed before rendering the complete scene. Although each adversary is optimized independently (without explicit coordination), later agents react to earlier ones, yielding complex multi-threat scenarios for stress-testing autonomous driving systems. Example scenes with multiple adversarial vehicles are shown in Figure 5, with more on our project page.

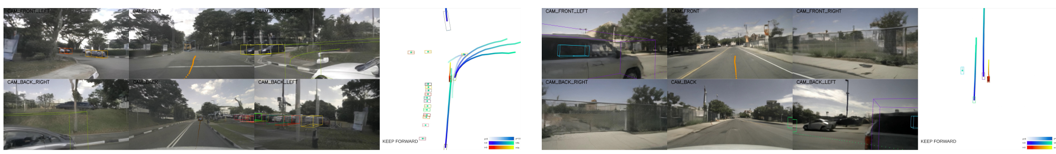
4 Experiments

4.1 Experiment Setup

Datasets. We adopt the widely used nuScenes dataset [7] as the foundation for our proposed *Challenger* framework, enabling seamless integration with existing end-to-end autonomous driving (AD) models. It is a large-scale dataset containing 1,000 scenes where each scene is 20 seconds long. Each scene is annotated with 3D bounding boxes of objects, including vehicles, pedestrians, and obstacles, at a frequency of 2Hz, and each annotated frame can serve as a sample for an E2E AD system. Specifically, we use the nuScenes validation set, which contains 150 scenes (6,019 samples).

Dataset	AD Model	1s	2s	3s	Avg	Degradation
nuScenes-val	UniAD [23]	0.10%	0.15%	0.61%	0.29%	
nuScenes-val-R		0.07%	0.24%	0.64%	0.32%	
Adv-nuSc		**0.80%**	**4.10%**	**6.96%**	**3.95%**	12.6× ↑
nuScenes-val	VAD [27]	0.11%	0.24%	0.42%	0.26%	
nuScenes-val-R		0.19%	0.32%	0.59%	0.37%	
Adv-nuSc		**4.46%**	**7.59%**	**9.08%**	**7.05%**	26.1× ↑
nuScenes-val	SparseDrive [45]	0.020%	0.063%	0.238%	0.107%	
nuScenes-val-R		0.010%	0.090%	0.370%	0.156%	
Adv-nuSc		**0.029%**	**0.618%**	**2.430%**	**1.026%**	8.6× ↑
nuScenes-val	DiffusionDrive [33]	0.059%	0.068%	0.169%	0.099%	
nuScenes-val-R		0.030%	0.055%	0.403%	0.163%	
Adv-nuSc		**0.068%**	**1.299%**	**3.646%**	**1.671%**	15.9× ↑

Table 1: **Performance of End-to-End Autonomous Driving Systems Across Different Datasets.** We report the collision rate over 3 seconds and the average collision rate for each dataset. Tested E2E AD models perform much worse on adversarial driving scenarios generated by *Challenger*.



(a) Surpassing from the right. (b) Cut-in from the left.

Figure 4: Failure cases of UniAD in adversarial scenarios generated by *Challenger*.

To investigate potential distributional differences between the original camera data and our rendered outputs, we additionally re-render the original nuScenes validation set using our multiview neural renderer. We refer to this rendered version as *nuScenes-val-R*. To evaluate adversarial robustness, we generate a new dataset, *Adv-nuSc*, using *Challenger*. It contains 156 scenes (6,115 samples) and is specifically crafted to challenge the ego vehicle by introducing adversarial traffic participants.

Autonomous Driving Models. We evaluate four state-of-the-art vision-based end-to-end AD models: UniAD [23], VAD [27], SparseDrive [45], and DiffusionDrive [33]. For each method, we use official pretrained models and use the default settings for evaluation. All models take multiview camera images as input and predict the ego vehicle's motion trajectory for the next 3 seconds. This unified setup allows for a fair and direct comparison across models on both standard and adversarial test sets.

4.2 Results on the *Adv-nuSc* Dataset

We evaluate the performance of four end-to-end autonomous driving systems on three datasets: the original nuScenes validation set, the re-rendered *nuScenes-val-R*, and our *Adv-nuSc* dataset. Results are summarized in Table 1. We report the collision rate over 3 seconds—defined as the proportion of evaluated samples in which the planned trajectory collides with any object.

Successfully challenging E2E AD algorithms. As shown in Table 1, all models perform comparably on the original nuScenes dataset and its re-rendered counterpart *nuScenes-val-R* with limited degradation, indicating that distribution shifts between these datasets do exist but are minimal. In contrast, all models exhibit a substantial increase in collision rates on the *Adv-nuSc* dataset. These results demonstrate that *Challenger* produces adversarial scenarios that serve as effective safety test cases, exposing the failure modes of existing autonomous driving systems, and revealing critical gaps in their robustness and generalization. We provide qualitative examples in Section 4.3 to analyze how these adversarial driving scenarios challenge E2E AD systems under safety-focused testing.

4.3 Qualitative Examples of Adversarial Driving Scenarios

We present two representative adversarial scenarios generated by *Challenger* and analyze the behavior of a representative E2E AD model, UniAD, as shown in Figure 4.

(a) Being sandwiched.

(b) Double overtaking.

Figure 5: *Challenger* can generate scenarios with multiple adversarial vehicles (white 3D bounding boxes).

MTR	TS	Collision Rate
✓	✓	**6.17%**
×	✓	3.49%
✓	×	0.13%
×	×	0.80%

Table 2: **Ablation study.** We show the effectiveness of the multi-round trajectory refinement process (MTR) and trajectory scoring process (TS).

Dataset	SC	IQ
nuScenes-val	0.8349	0.3801
nuScenes-val-R	0.8173	0.3471
Adv-nuSc	0.8093	0.3469

Table 3: **Quantitative evaluation of video quality.** We report the subject consistency (SC) and imaging quality (IQ) of the datasets.

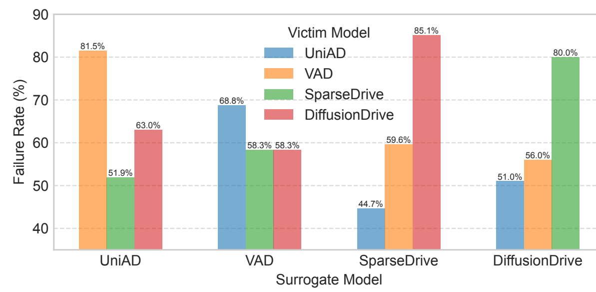


Figure 6: **Attack Transferability.** This figure illustrates the failure rates of victim models when evaluated on adversarial driving scenarios filtered based on the performance of a surrogate model. A failure is recorded when the victim model collides with another object in at least one sample of a given scenario. The failure rate is computed as the number of such scenarios divided by the total number of filtered scenarios.

In Figure 4a, the adversarial vehicle—outlined with a green bounding box—is positioned to the right of the ego vehicle. Its actual intention is to surpass the ego vehicle from the right lane. However, the UniAD model incorrectly predicts that the vehicle intends to turn right and thus fails to yield. This misjudgment likely stems from dataset bias, as vehicles in the right lane typically perform right turns during training, leading the model to generalize incorrectly.

In Figure 4b, the adversarial vehicle—highlighted in a purple bounding box—initiates a sudden cut-in from the left lane. The UniAD model, expecting the vehicle to maintain a straight path, continues forward without adjusting its trajectory. This behavior reflects the rarity of such aggressive maneuvers in training data, limiting the model's generalization to out-of-distribution scenarios.

These examples show how *Challenger* tests E2E AD safety and robustness under challenging conditions, revealing vulnerabilities hidden in naturalistic testing.

4.4 Attack Transferability

Curating transferable test cases. We further show that the proposed *Challenger* framework can be reformulated to perform transferable adversarial attacks on E2E AD models with significant success rates. Concretely, we integrate a surrogate model (e.g., UniAD) into the pipeline, generate a pool of adversarial driving scenes and filter them by only keeping those in which the surrogate model fails to navigate safely in at least one sample (i.e., only retain scenarios where UniAD fails). This filtering strategy is designed to select scenarios that are sufficiently challenging and thus more likely to induce failure in other AD models. We subsequently evaluate the remaining victim models on this curated subset. An attack is considered *transferable* if a victim model also experiences a collision in at least one sample of a scenario that was selected based on surrogate failure.

Observed transferable safety failures. To quantify this effect, we conduct pairwise evaluations across the four tested E2E AD models, using each model in turn as a surrogate and assessing the failure rate of the remaining three as victims. Results, shown in Figure 6, reveal that a substantial

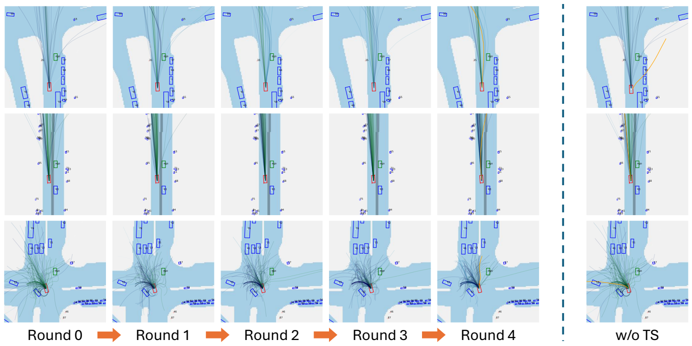


Figure 7: **Qualitative example of the multi-round trajectory refinement process and trajectory scoring**. Multi-round trajectory refinement helps the candidate trajectories converge to better trajectories while trajectory scoring effectively gathers context information. Please refer to Figure 1 for a legend.

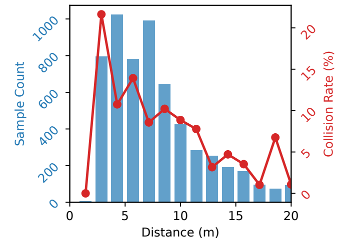


Figure 8: Correlation between distance and collision rate. Tested on the Adv-nuSc dataset with the UniAD model.

Figure 9: **Representative frames from datasets.** The comparison illustrates that the re-rendered and adversarial datasets maintain photorealistic quality comparable to the original nuScenes dataset.

portion of the filtered scenarios lead to failure in models other than the one used for filtering. These findings highlight *Challenger*'s ability to generate stressful test cases and raise critical safety implications, suggesting that current E2E AD systems may share vulnerabilities and remain susceptible to transferable adversarial attacks.

4.5 Ablation Study.

To assess the contributions of the multi-round trajectory refinement and trajectory scoring processes within *Challenger*, we conduct an ablation study. We randomly select 20 scenes from the nuScenes validation set as initializations and run *Challenger* under various configurations. For each configuration, we report the percentage of samples that *Challenger* could successfully generate while the representative UniAD model failed to navigate safely. The results are summarized in Table 2. Additionally, we provide qualitative examples in Figure 7. Intuitively, with multi-round trajectory refinement, the candidate trajectories gradually converge to better trajectories. Without trajectory scoring, context information cannot be effectively gathered, and the generated trajectory is not challenging. Additionally, we analyzed the correlation between the distance to the adversarial vehicle and the collision rate in Figure 8, validating our scoring function design that rewards close encounters.

4.6 Video Quality Evaluation

To evaluate dataset visual realism, we conduct qualitative and quantitative assessments. Representative examples are provided in Figure 9, covering daytime, evening, and rainy weather. For quantitative evaluation, we adopt two metrics: *subject consistency* and *imaging quality*, and report results in Table 3. The protocol follows a two-stage adaptation of VBench for multi-object driving scenes.

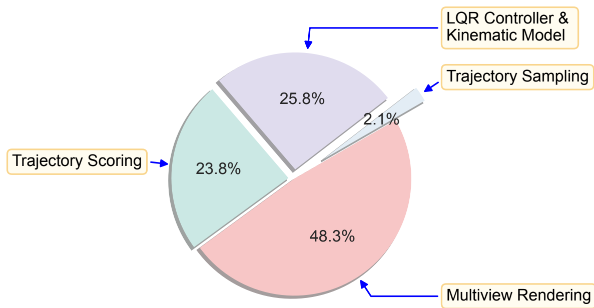


Figure 10: **Runtime Composition.** Percentage of time spent on each major *Challenger* component.

Component	Runtime (s)
Traj. Sampling	17.36
Physics	211.36
Traj. Scoring	194.72
Rendering	395.18
Total	818.62

Table 4: **Runtime Breakdown.** Time spent on trajectory sampling, physics simulation (including the LQR controller and kinematic model), scoring, and multiview neural rendering.

Comparable photorealistic quality. Our findings show that both the re-rendered *nuScenes-val-R* and adversarial *Adv-nuSc* datasets have photorealistic quality generally comparable to original nuScenes, qualitatively and quantitatively. This indicates that *Challenger* provides a high-fidelity testing environment for evaluating E2E AD safety and robustness.

4.7 Runtime Analysis

We provide a runtime breakdown of *Challenger*, highlighting the computational cost of each component for one adversarial driving scene. All experiments use a cloud server with dual Intel(R) Xeon(R) Platinum 8457C CPUs, 4 NVIDIA L20 GPUs (48GB each), and 512GB RAM. During adversarial trajectory generation, one GPU samples trajectories from the diffusion model, while physics-based simulation and trajectory scoring run primarily on the CPU. Multiview neural rendering is distributed across all four GPUs.

Affordable cost, with room for optimization. As shown in Figure 10 and Table 4, multiview neural rendering dominates runtime (about half) because it synthesizes high-resolution, six-view videos over hundreds of frames. This underscores the efficiency of our adversarial trajectory generation design: multi-round refinement and scoring enable *Challenger* to generate high-quality adversarial plans without rendering every candidate trajectory, which would be prohibitive. Physics simulation and trajectory scoring are also non-trivial due to batched, multi-round evaluation, but remain cost-effective versus full rendering. Current CPU-bound implementations could be further accelerated with GPU or parallelization.

5 Conclusion and Future Work

Safety-critical systems require testing beyond average-case performance, and we introduced *Challenger*, a framework that generates *physically plausible, photorealistic, and challenging yet solvable* driving videos to probe safety margins of end-to-end autonomous driving (E2E AD) models. By combining a diffusion-based trajectory generator, a physics-aware planning simulator, and a multiview neural renderer, *Challenger* provides controlled adversarial exposures that reveal safety-relevant failure modes and transfer across state-of-the-art models. A key limitation is the fully open-loop evaluation: models do not influence scenario evolution, which may deviate from real-world interaction dynamics. We view this setting as a conservative, audit-friendly stress test, but not a substitute for closed-loop and real-world safety validation; training on these videos alone cannot guarantee safe deployment. Future work will pursue closed-loop integration, multi-modality rendering, and alignment with certification practices and third-party audit protocols, with auto-regressive driving video generation as a path to generate future frames based on model actions. We aim for *Challenger* to ethically and rigorously pressure-test AD systems, contributing to safer and more trustworthy autonomy.

References

[1] Tesla Vision Update: Replacing Ultrasonic Sensors with Tesla Vision | Tesla Support United Kingdom.

[2] Yasasa Abeysirigoonawardena, Kevin Xie, Chuhan Chen, Salar Hosseini, Ruiting Chen, Ruiqi Wang, and Florian Shkurti. Generating transferable adversarial simulation scenarios for self-driving via neural rendering. *arXiv preprint arXiv:2309.15770*, 2023.

[3] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10017–10029, 2023.

[4] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17890–17901, 2024.

[5] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[6] Adith Boloor, Karthik Garimella, Xin He, Christopher Gill, Yevgeniy Vorobeychik, and Xuan Zhang. Attacking vision-based perception in end-to-end autonomous driving models. *Journal of Systems Architecture*, 110:101766, 2020.

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[8] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. (arXiv:2106.11810), February 2022. arXiv:2106.11810 [cs].

[9] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating, 2019.

[10] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15793–15803, 2021.

[11] Sehwan Choi, Jungho Kim, Junyong Yun, and Jun Won Choi. R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8525–8535, 2023.

[12] OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. `https://github.com/OpenDriveLab/OpenScene`, 2023.

[13] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, December 2024.

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, page 8780–8794. Curran Associates, Inc., 2021.

[15] Wenhao Ding, Baiming Chen, Minjun Xu, and Ding Zhao. Learning to collide: An adaptive safety-critical scenarios generating method. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 2243–2250, October 2020.

[16] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953):620–627, 2023.

[17] Shuo Feng, Xintao Yan, Haowei Sun, Yiheng Feng, and Henry X Liu. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature communications*, 12(1):748, 2021.

[18] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024.

[19] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrivedit: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024.

[20] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *ICLR*, 2024.

[21] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2025.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, page 6840–6851. Curran Associates, Inc., 2020.

[23] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023.

[24] Zhenbo Huang, Shiliang Sun, Jing Zhao, and Liang Mao. Multi-modal policy fusion for end-to-end autonomous driving. *Information Fusion*, 98:101834, 2023.

[25] Neil Jacobstein. Autonomous vehicles: An imperfect path to saving millions of lives. *Science Robotics*, 4(28):eaaw8703, March 2019. Publisher: American Association for the Advancement of Science.

[26] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. November 2024.

[27] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 8306–8316, October 2023.

[28] ByeoungDo Kim, Chang Mook Kang, Seung Hi Lee, Hyunmin Chae, Jaekyum Kim, Chung Choo Chung, and Jun Won Choi. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network, 2017.

[29] Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, pages 131–148. Springer, 2024.

[30] Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. *arXiv preprint arXiv:2412.05435*, 2024.

[31] Bohan Li, Yasheng Sun, Zhujin Liang, Dalong Du, Zhuanghui Zhang, Xiaofeng Wang, Yunnan Wang, Xin Jin, and Wenjun Zeng. Bridging stereo geometry and bev representation with reliable mutual interaction for semantic scene completion. *arXiv preprint arXiv:2303.13959*, 2023.

[32] Shuyang Li, Talha Azfar, and Ruimin Ke. Chatsumo: Large language model for automating traffic scenario generation in simulation of urban mobility. *IEEE Transactions on Intelligent Vehicles*, 2024.

[33] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. (arXiv:2411.15139), November 2024. arXiv:2411.15139 [cs].

[34] Henry X. Liu and Shuo Feng. Curse of rarity for autonomous vehicles. *Nature Communications*, 15(1):4808, June 2024. Publisher: Nature Publishing Group.

[35] William Ljungbergh, Adam Tonderski, Joakim Johnander, Holger Caesar, Kalle Åström, Michael Felsberg, and Christoffer Petersson. *NeuroNCAP: Photorealistic Closed-Loop Safety Testing for Autonomous Driving*, volume 15088 of *Lecture Notes in Computer Science*, page 161–177. Springer Nature Switzerland, Cham, 2025.

[36] Jiageng Mao, Boyi Li, Boris Ivanovic, Yuxiao Chen, Yan Wang, Yurong You, Chaowei Xiao, Danfei Xu, Marco Pavone, and Yue Wang. Dreamdrive: Generative 4d scene modeling from street view images. *arXiv preprint arXiv:2501.00601*, 2024.

[37] Jeffrey Mervis. Not so fast. *Science*, December 2017. Publisher: American Association for the Advancement of Science.

[38] Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. Evaluating uncertainty quantification in end-to-end autonomous driving control. *arXiv preprint arXiv:1811.06817*, 2018.

[39] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024.

[40] Nick Petrić Howe and Shamini Bundell. How to make driverless cars safer — expose them to lots of dangerous drivers. *Nature*, March 2023.

[41] Philip Polack, Florent Altché, Brigitte d'Andréa Novel, and Arnaud de La Fortelle. The kinematic bicycle model: A consistent model for planning feasible trajectories for autonomous vehicles? In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 812–818, 2017.

[42] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7077–7087, 2021.

[43] Jay Sarva, Jingkang Wang, James Tu, Yuwen Xiong, Sivabalan Manivasagam, and Raquel Urtasun. Adv3d: Generating safety-critical 3d objects through closed-loop simulation. *arXiv preprint arXiv:2311.01446*, 2023.

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. October 2020.

[45] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. (arXiv:2405.19620), May 2024. arXiv:2405.19620.

[46] Shuhan Tan, Boris Ivanovic, Xinshuo Weng, Marco Pavone, and Philipp Kraehenbuehl. Language conditioned traffic generation. *arXiv preprint arXiv:2307.07947*, 2023.

[47] James Tu, Simon Suo, Chris Zhang, Kelvin Wong, and Raquel Urtasun. Towards scalable coverage-based testing of autonomous vehicles. In *Conference on Robot Learning*, pages 2611–2623. PMLR, 2023.

[48] Lening Wang, Wenzhao Zheng, Dalong Du, Yunpeng Zhang, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, Jie Zhou, Jiwen Lu, et al. Stag-1: Towards realistic 4d driving simulation with video generation model. *arXiv preprint arXiv:2412.05280*, 2024.

[49] Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. *arXiv preprint arXiv:2410.18079*, 2024.

[50] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *ECCV*, 2024.

[51] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 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*, pages 14749–14759, 2024.

[52] Junkai Xia, Chenxin Xu, Qingyao Xu, Yanfeng Wang, and Siheng Chen. Language-driven interactive traffic trajectory generation. *Advances in Neural Information Processing Systems*, 37:77831–77859, 2024.

[53] Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemma: Open-source multimodal model for end-to-end autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1001–1009, 2025.

[54] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.

[55] Xintao Yan, Zhengxia Zou, Shuo Feng, Haojie Zhu, Haowei Sun, and Henry X Liu. Learning naturalistic driving environment with statistical realism. *Nature communications*, 14(1):2037, 2023.

[56] Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, et al. Streetcrafter: Street view synthesis with controllable video diffusion models. *arXiv preprint arXiv:2412.13188*, 2024.

[57] Brian Yang, Huangyuan Su, Nikolaos Gkanatsios, Tsung-Wei Ke, Ayush Jain, Jeff Schneider, and Katerina Fragkiadaki. Diffusion-es: Gradient-free planning with diffusion for autonomous and instruction-guided driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15342–15353, 2024.

[58] Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-end autonomous driving through v2x cooperation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9598–9606, 2025.

[59] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024.

[60] Linrui Zhang, Zhenghao Peng, Quanyi Li, and Bolei Zhou. Cat: Closed-loop adversarial training for safe end-to-end driving. In *Conference on Robot Learning*, pages 2357–2372. PMLR, 2023.

[61] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

[62] Wenzhao Zheng, Ruiqi Song, Xianda Guo, and Long Chen. Genad: Generative end-to-end autonomous driving. *arXiv preprint arXiv:2402.11502*, 2024.

[63] Yinan Zheng, Ruiming Liang, Kexin Zheng, Jinliang Zheng, Liyuan Mao, Jianxiong Li, Weihao Gu, Rui Ai, Shengbo Eben Li, Xianyuan Zhan, et al. Diffusion-based planning for autonomous driving with flexible guidance. *arXiv preprint arXiv:2501.15564*, 2025.