Gen-NCAP: A Generative Simulator for Corner Case Benchmarking in End-to-End Autonomous Driving

Gen Li[1,2], Nan Wang[3], YunLong Li[1], Ziyang Yan[3]
Yinghao Shuai[7], Saining Zhang[8], Shu Han[3], Shaocong Xu[3]
Lu Zhang[4],Luan Zhang[4],Paul Barsch[4],Zebang Shen[4]
Hang Zhao[6],Yaqin Zhang[1],Chaojian Li[5,†],Hao Zhao[1,3]
[1]AIR, Tsinghua University, [2]Zhejiang University, [3]Beijing Academy of Artificial Intelligence
[4]Mercedes-Benz, [5]The Hong Kong University of Science and Technology
[6]IIIS, Tsinghua University, [7]Tongji University, [8]Nanyang Technological University
[†]Corresponding author.

Abstract

Recent advances in end-to-end autonomous driving (E2E-AD), enabled by foundation models have shown great promise towards full self-driving. To fairly evaluate E2E-AD frameworks, simulation-based benchmarks are used to test performance in safety-critical scenarios, offering a more convenient and controlled alternative to on-road testing. However, their effectiveness is limited by (1) low extensibility, due to manual effort in scenario design, and (2) a large sim-to-real gap, caused by low photorealism, both especially problematic for rare corner cases. To address these issues, inspired by the safety-critical corner case design in the European New Car Assessment Programme (Euro NCAP), we propose **Gen-NCAP**, to the best of our knowledge, the first generative closed-loop simulator for benchmarking E2E-AD corner cases. Gen-NCAP improves extensibility by combining (1) a vision-language model that places them appropriately in a scene, enabling scenario creation using only text prompts while enhancing coverage by supporting multiple corner cases and (2) a 3D object generation model that enhances flexibility by creating diverse and high-quality objects from text. To narrow the sim-to-real gap, we introduce a video refinement pipeline using world models trained on datasets curated to (1) fix artifacts in closed-loop simulation, i.e., foreground-background misalignment arising from the insertion of complete 3D objects into existing scenes, and blurriness in novel view synthesis, thereby improving visual fidelity, and (2) ensure temporal consistency of the refinement results across frames to maintain the stability of the simulator. Using Gen-NCAP, we release a benchmark of 300 interactive scenarios across 8 types of corner cases and introduce a fine-grained collision metric for closed-loop evaluation. With the proposed Gen-NCAP, E2E-AD developers and researchers can not only leverage the generated data for open-loop training to enhance model robustness in safety-critical situations but also conduct fair and reproducible benchmarking of different E2E-AD frameworks.

1 Introduction

Recent breakthroughs in autonomous driving have demonstrated the effectiveness of end-to-end autonomous driving (E2E-AD) models that directly output planned trajectories from sensor inputs [1, 2, 3, 4, 5]. Despite the great promise of fully self-driving vehicles enabled by data-driven E2E-AD, evaluating such systems remains a significant challenge [6, 7, 8, 9, 10]. Specifically, direct on-road testing is costly and does not enable fair benchmarking across different E2E-AD models. Meanwhile,

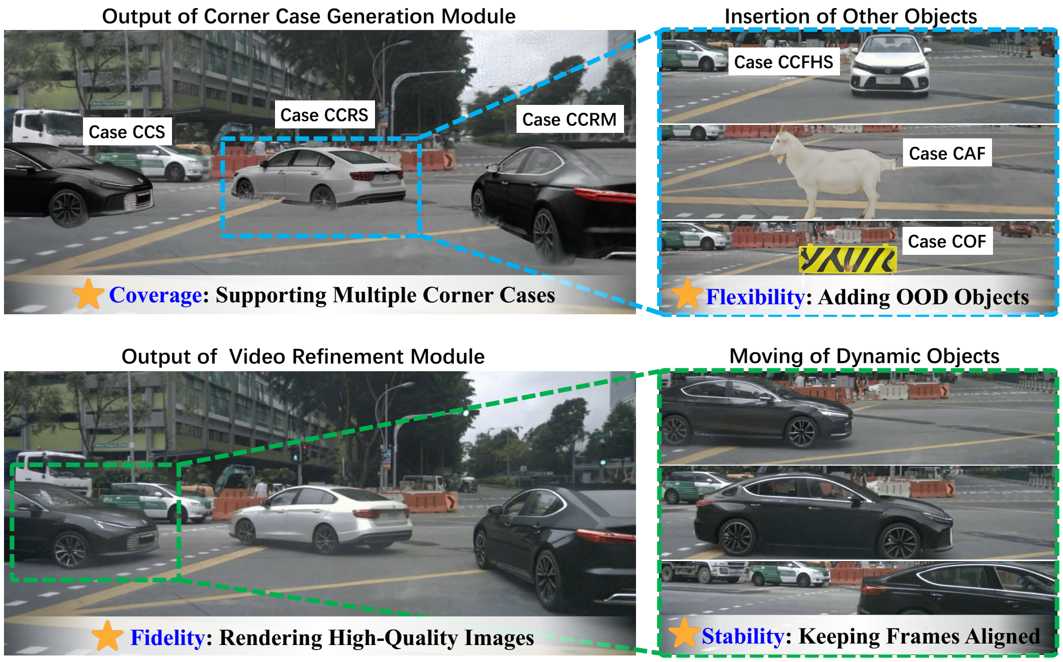


Figure 1: We propose **Gen-NCAP**, a generative closed-loop simulator demonstrating strong extensibility across diverse corner cases and reduced sim-to-real gap through high-fidelity, temporally consistent video refinement module.

log-replay of recorded trajectories only provides open-loop evaluation, lacking the ability to interact with the surrounding environment based on the model's outputs. As a result, simulation-based closed-loop evaluation [9, 10, 11] has gained increasing attention as a convenient and controllable approach, especially for benchmarking safety-critical corner cases where timely and accurate environmental feedback is essential. However, the effectiveness of existing simulation-based benchmarks is hindered by two key challenges. (1) **Low extensibility**: Although environmental feedback can be automatically obtained from simulators, designing each corner case still requires substantial manual effort, and the resulting scenarios often exhibit limited diversity and low quality in the objects involved. This issue can be further categorized into two dimensions, namely coverage and flexibility, as illustrated below.

Identified Challenge 1A: Coverage

The traditional manual design of corner cases requires substantial manual effort, which limits the volume of scenarios in closed-loop simulation benchmarks.

Identified Challenge 1B: Flexibility

Existing simulation-based benchmarks can only augment previously seen scenes with known objects and lack the capability to generate diverse, high-quality ones. Moreover, most corner cases inherently involve out-of-distribution (OOD) objects.

(2) **Large sim-to-real gap:** In closed-loop simulation, the sim2real gap, defined as the discrepancy between simulated and real-world dynamics, affects model reliability and comprises two key aspects, fidelity and stability, as illustrated below.

Identified Challenge 2A: Fidelity

In simulation environments, directly inserting 3D objects into existing scenes can cause foreground-background misalignment, and imperfect 3D reconstruction may introduce artifacts in novel-view rendering, posing critical challenges for reliable benchmarking.

Identified Challenge 2B: Stability

Existing world models for 3D reconstruction artifact restoration overlook dynamic foregrounds and lack temporal consistency, causing instability in closed-loop simulation.

To address these challenges, we propose **Gen-NCAP**, to the best of our knowledge, the first generative closed-loop simulator for benchmarking E2E-AD corner cases that achieves high rendering quality, large-scale scene generation, and strong extensibility (see Fig. 1). Inspired by the corner case design principles in the European New Car Assessment Programme (Euro NCAP) [12], Gen-NCAP provides a scalable and reproducible framework for such evaluations. To tackle the **extensibility** issue, Gen-NCAP integrates a generative pipeline combining a 3D object generation model (to synthesize cars, pedestrians, etc.) and a vision-language model (to plan their layout, including size and placement) based on natural language prompts, eliminating the need for manual scene design. To bridge the **sim-to-real gap**, we train a unified video diffusion model on a carefully curated dataset to correct foreground-background misalignment resulting from the insertion of 3D vehicles into existing scenes, as well as blurriness in novel-view synthesis. Moreover, to ensure the stability of closed-loop simulation, the trained video diffusion model is applied to the simulator to preserve temporal consistency. This refinement is agnostic to the underlying scene representation used by the simulator. Built on Gen-NCAP, we curate a benchmark comprising 300 interactive scenarios across eight categories of corner cases and introduce a dedicated metric to assess E2E-AD model performance on this benchmark.

Built on top of the contributions illustrated above, Gen-NCAP offers the following features:

- **Generative Scenario Design:** Enables scalable creation of diverse, safety-critical driving scenes using text prompts, enhancing extensibility for large-scale evaluations.
- **Photorealism via Video Refinement:** Minimizes the sim-to-real gap through a novel video refinement pipeline, reducing the FID by **29.49**.
- **Fine-Grained Closed-Loop Benchmarking:** Provides **300** interactive scenarios and a collision-based metric for detailed E2E-AD evaluation.
- **Support for Open-Loop Training:** Generated scenes can augment datasets, improving E2E-AD robustness on corner cases. Fine-tuning VAD with Gen-NCAP data reduces collision rate by **5%** on the NeuroNcap benchmark.

These capabilities make Gen-NCAP a novel and practical tool for researchers aiming to evaluate and improve the safety and robustness of modern E2E-AD systems in rare but critical driving scenarios.

2 Related Works

2.1 End-to-End Autonomous Driving

The development of end-to-end autonomous driving systems [1, 2, 3, 4, 13, 14, 5] allows models to map raw sensor data directly to trajectories or control commands, eliminating the need for manually engineered features. This approach enhances data efficiency and improves adaptability across diverse driving conditions [15]. For instance, SparseDrive [16] employs symmetric sparse perception to support parallel motion planning, but its simplistic architecture and lack of ego-centric interaction constrain both planning performance and computational efficiency. DiffusionDrive [17] proposes a truncated diffusion policy that denoises an anchored Gaussian distribution into a multi-modal distribution of driving actions while UniAD [18] introduces a goal-conditioned planning framework, which integrates perception and motion prediction modules to improve decision-making. Similarly, VAD [19] improves end-to-end planning through the use of vectorized scene representations.

2.2 Evaluation Benchmarks for E2E-AD

Benchmarking is crucial for E2E-AD research, providing standardized metrics for performance and practicality. Closed-loop evaluation captures interactions and metrics like collision rate and rule compliance. CARLA [20] supports several benchmarks but relies on synthetic scenarios. Platforms like nuPlan [21] and Waymax [22] use realistic trajectories but lack full end-to-end pipelines. NeRF-based simulators such as NeuroNCAP [9] and Unisim [11] offer photorealistic sensor data but face

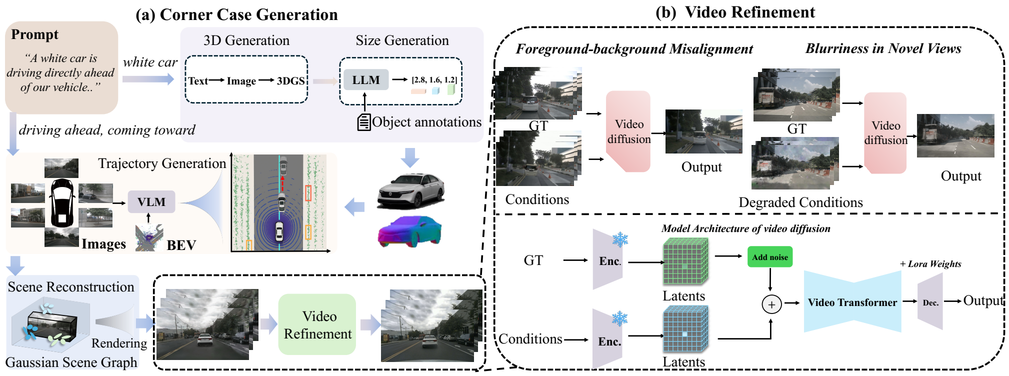


Figure 2: Overview of Gen-NCAP. Our Gen-NCAP framework consists of two key modules: (a) a Corner Case Generation Module, which generates a diverse range of safety-critical driving scenarios, and (b) a Video Refinement Module, which reduces rendering artifacts using targeted video diffusion models, minimizing the sim-to-real gap.

scalability challenges [23, 24, 25], and current methods struggle with complex trajectories. Gen-NCAP overcomes these limitations with a generative closed-loop benchmark using real-world traffic data, providing high-quality and highly extensible evaluation for E2E-AD systems.

2.3 Generative Model for AD Simulation

Novel-view rendering quality remains a key challenge in autonomous driving simulation. Recent generative models enhance realism and scalability [26]. FreeVS [27] uses pseudo-image view priors with viewpoint transformation to emulate camera motion, while FreeSim [28] combines a generative enhancement model with progressive reconstruction for off-trajectory views. StreetCrafter [29] leverages LiDAR renderings and generative priors, ReconDreamer [30] applies online refinement, and Difix3D+ [31] uses a one-step diffusion model with progressive updates. Despite progress, these methods neglect foreground refinement and 4D temporal consistency.

3 Gen-NCAP Simulator

3.1 Gen-NCAP: Framework Overview

In this section, we introduce Gen-NCAP, the first generative closed-loop simulator for benchmarking corner cases. As illustrated in Fig. 2. Gen-NCAP comprises two core components: (1) Corner Case Generation, which produces high-quality, highly controllable autonomous driving corner case scenarios with low sim2real gaps based on text prompts inputs and (2) a Video Refinement module, which delivers high-fidelity rendering for closed-loop simulation.

3.2 Gen-NCAP: Corner Cases Generation

Object Generation. To achieve high-quality and strong extensibility closed-loop Simulator, we design a text-to-image-to-3DGS object generation pipeline. Specifically, we use text prompts to generate high-quality images via FLUX.1-dev [32], and then feed these images into TRELLIS [33], a SoTA 3D generation method, to create high-quality 3D Gaussian vehicle assets G_{obj}. Importantly, since assets generated or retrieved often lack physically plausible sizes [34], we leverage GPT-4o's natural language understanding capabilities to learn typical object dimensions from annotated categories in the driving scene. GPT-4 then outputs the most contextually appropriate size for each object. The generated results of the objects are shown in Fig. 3(a).

Technique 1A: Object Generation Module

We propose a text-to-3D object generation module to produce high-quality and diverse objects, addressing the flexibility issue in corner case simulation.

(a) Object Generation

Text Prompt：*A white car is moving across the front-right area relative to our ego vehicle.*

(b) Trajectory Generation

Figure 3: Visualizations of generated objects and trajectories.

Trajectory Generation. To construct moving foreground objects, we leverage the powerful spatial understanding capabilities of GPT-4o to generate realistic motion trajectories for the objects. Specifically, we use the ego vehicle as the origin of the coordinate system to build the scene's bird's-eye view (BEV) representation. This includes the BEV coordinates and bounding box dimensions of the existing objects. Then, we combine the spatial information including BEV information and image with target object's attributes including semantic category and scale parameters to feed into the VLM allowing VLM to understand the spatial representation at the current time. We then provide VLM with the desired Cornercase prompt, generating a series of trajectories in the BEV coordinate system $p_t^{\mathrm{BEV}} = (x_t, y_t) \in \mathbb{R}^2$.These trajectories are subsequently transformed into the world coordinate system using the ego-to-world transformation matrix $\tilde{\mathbf{p}}_t^{\mathrm{BEV}} = [x_t, y_t, 0, 1]^\top \in \mathbb{R}^4$, $\tilde{\mathbf{p}}_t^{\mathrm{world}} = \mathbf{T}_{\mathrm{ego2world}} \cdot \tilde{\mathbf{p}}_t^{\mathrm{BEV}}$ yielding the object's global trajectory, where $\tilde{\mathbf{p}}_t^{\mathrm{world}} = [x_t^w, y_t^w, z_t^w, 1]^\top$. The trajectory generation results are shown in Fig. 3(b).

Technique 1B: Trajectory Generation Module

We propose a vision-language-based trajectory generation module that uses BEV and image inputs to generate object positions and trajectories from text prompts, enabling Gen-NCAP to efficiently design large-scale corner case scenarios and cover a wide variety of corner case types.

Driving Scene Reconstruction. To simulate driving scenarios, we employ 3D Gaussian Splatting [35] to reconstruct a 3D environment from real-world data, allowing sensor rendering from arbitrary viewpoints and enabling closed-loop simulation. To support flexible scene editing, we adopt the Gaussian Scene Graph [36], where the scene is represented by distinct nodes: sky nodes $\mathcal{G}^{sky}$, background nodes $\mathcal{G}^{bg}$, rigid nodes $\mathcal{G}^{rigid}$ (e.g., vehicles), and non-rigid nodes $\mathcal{G}^{non-rigid}$ (e.g., pedestrians). This structured representation allows independent editing of each Gaussian node while preserving reconstruction quality, forming a semantically organized 3D driving scene.

4 Gen-NCAP Simulator

4.1 Gen-NCAP: Video Refinement Module

To enable high-quality closed-loop simulation and narrow the sim-to-real gap, we propose a novel video refinement pipeline for driving scenario refinement. Our Video Refinement approach allows the world model [37] to learn the transition from low-quality to high-quality distributions.

Data Curation. To supervise our video diffusion model, we construct a large dataset of paired images with two types of artifacts and their corresponding ground-truth. The first type arises from foreground-background misalignment. To create these examples, we segment foreground vehicles from the nuScenes dataset and process them with TRELLIS [33] to obtain visually consistent 3D Gaussian rigid nodes. Replacing the original nodes with these Gaussian representations yields a low-quality foreground dataset that deliberately introduces misalignment. The second type of artifact appears in novel view synthesis. We simulate this by using under-trained driving scene reconstruction models to render low-quality videos, generating image pairs that exhibit perspective inconsistencies and distortions characteristic of imperfect novel view generation. Consequently, the constructed

Figure 4: Visualization of the paired dataset. (a) simulates artifacts in novel-view rendering, (b) simulates foreground–background misalignment.

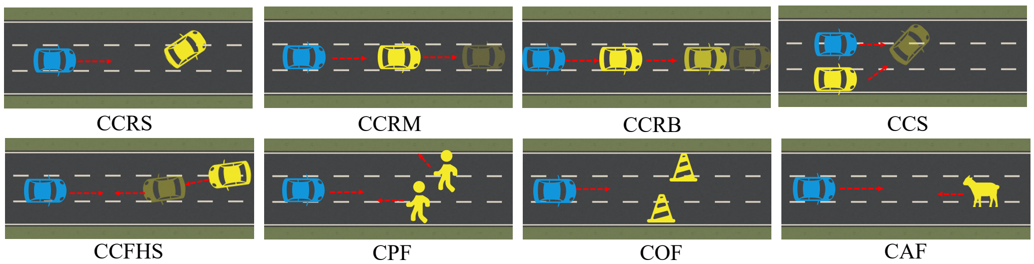


Figure 5: Illustration of the eight distinct categories of safety-critical scenarios in Gen-NCAP.

rendering data pairs are represented as $\left\{V^{'}, V_{\text{gt}}\right\}$. The dataset visualization is shown in Fig. 4.

Training of the Video Refinement Module. Once the above dataset has been constructed, we proceed to train the video diffusion model using these image pairs. We use $V^{'}$ as a control condition to provide appearance priors. The video transformer $\mathcal{F}_\theta$ is optimized by minimizing the following loss objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_i,\epsilon}\left[\left\|\mathbf{z}_i - \mathcal{F}_\theta\left(\bar{\mathbf{z}}_{i,t}, t\right)\right\|_2^2\right], \tag{1}$$

where $\mathbf{z}_i$ and $\hat{\mathbf{z}}_i$ are the latent codes of V_{gt} and $V_I^{'}$, respectively, obtained through a pre-trained VAE encoder.

Technique 2A: Training of the Video diffusion model

To jointly repair artifacts in both the foreground(vehicle) and background, we train a video refinement module on a large dataset consisting of paired clean and artifact-contaminated samples.

Application of the Video Refinement Module in Closed-Loop Simulation. Since the end-to-end closed-loop simulator can only access images from past time steps $\{i_0, i_1, ..., i_{t-1}\}$ and the current frame i_t, we combine these images into a video V_t and feed it into the trained video refinement model to correct artifacts and ensure temporal consistency. The refined video $V_t^{'}$ provides the artifact-free current frame $i_t^{'}$, which is then used as input to the AD model. More details of the closed-loop simulator are provided in the Appendix.

Technique 2B: Application of the Video Refinement Module in Closed-Loop Simulation

We adapt the video diffusion model to the characteristics of the end-to-end closed-loop simulator, ensuring temporal consistency while refining artifacts and maintaining the stability of the closed-loop simulation.

5 Gen-NCAP Benchmark

Using the generative simulator, we evaluate E2E-AD models in safety-critical scenarios. The Gen-NCAP dataset, derived from nuScenes [38], contains 300 driving scenarios, including 47 nuScenes clips and diverse Corner Case scenarios inspired by Euro NCAP protocols. These Corner Cases cover a wide range of rare but plausible events to test AD system robustness. Visualizations are shown in Fig. 5 and Fig. 6, and details of the eight scenario types are provided in the Appendix.

Figure 6: Visualization of dynamic corner cases at different timesteps generated by Gen-NCAP.

5.1 Gen-NCAP: Evaluation Metrics

In closed-loop simulation, traditional open-loop metrics such as L2 distance fail to reflect the true performance of E2E AD models. Therefore, we design two metrics to evaluate the effectiveness of E2E AD models in closed-loop scenarios: the first focuses on collision assessment, and the second evaluates the sensitivity of the ego vehicle's responses to corner cases.

Gen-NCAP Score(GNS). Traditional collision metrics such as Collision Rate can only determine whether a collision occurred, but in reality, the consequences of colliding with a pedestrian are very different from those of colliding with a road obstacle. Moreover, the impact of a high-speed collision differs significantly from that of a low-speed one. To accurately assess the performance of E2E-AD models, we design the following metric inspired by the concept of momentum in physics:

$$\text{GNS} = \begin{cases} 100 & \text{if no collision} \\ 100/(1 + k \cdot s \cdot v) & \text{otherwise} \end{cases}, \tag{2}$$

here, k is a weighting factor based on the category of the object involved in the collision, s represents the size (volume) of the collided object, and v is the speed of the ego vehicle at the time of collision.

Gen-NCAP Responsiveness Score(GNRS). To evaluate the AD model's responsiveness to Corner-Cases in closed-loop simulation, we propose a new metric based on the ego vehicle's acceleration and the L2 distance between the closed-loop trajectory and the reference trajectory annotated in nuScenes.

$$\text{GNRS} = \log(1 + a^2) \times (1 + L), \tag{3}$$

where a is acceleration and L represent L2 distance. A higher value of GNRS indicates greater sensitivity and responsiveness of the autonomous driving model to corner-case scenarios.

6 Experiments

Dataset. Among various autonomous driving datasets, nuScenes [38] is widely used for end-to-end planning, from which we selected 47 scenes to generate corner cases. Additionally, we curated an extra dataset of unprocessed novel-view images from closed-loop simulations for evaluating the video refinement module (details in the appendix).

E2E-AD Models. In our benchmark, we evaluate four representative SoTA end-to-end autonomous driving models: UniAD [18], VAD [19], SparseDrive [16], and DiffusionDrive [17]. For evaluation, we use the pre-trained weights provided by the authors. More detailed on evaluation of E2E-AD models can be found in the Appendix.

Metric. In closed-loop evaluation, conventional open-loop metrics such as L2 distance often fail to accurately reflect the performance of AD models. Therefore, we use Gen-NCAP Score(GNS) and Gen-NCAP Responsiveness Score(GNRS) with NeuroNcap Score(NNS) [9], Collision Rate(CR) in NeuroNCAP [9].In addition, Acceleration is selected to evaluate driving comfort. To evaluate the quality of images without GT, we use FID[39], FID-A(FID on vehicle-centric crops), and FVD[40].

6.1 Gen-NCAP Enables Fair Evaluation

The quantitative results of the AD models are shown in Tab. 1 we observe that DiffusionDrive achieves the lowest collision rate among the quantitative metrics, while VAD attains the highest

Figure 7: Qualitative comparisons of the failure cases of four E2E-AD models in handling corner cases within closed-loop simulation.

scores in NNS and GNS. UniAD, however, achieves the best performance in GNRS. Notably, as shown in Tab. 1, among the four AD models, SparseDrive performs the worst across all metrics. This is partly due to its sparse perception design, which unifies detection, tracking, and mapping but tends to fall into the **"lazy regime"** [41], limiting its generalization in complex scenarios. Furthermore, the results also suggest that, while ride comfort is often emphasized in closed-loop evaluations, our corner case analysis reveals a trade-off, reducing acceleration improves comfort but increases collision risk and degrades NNS and GDS performance. This implies that **a certain level of discomfort may be necessary to ensure safety**. The qualitative results in Fig. 7 illustrate the failure cases of the four AD models in the corner cases we designed.

Table 1: Quantitative results of four E2E-AD models on the Gen-NCAP benchmark. Horizontal bars provide an intuitive comparison of performance across different metrics.

Metrics	Method	CCRS	CCRM	CCRB	CCS	CCFHS	CPF	COF	CAF	Avg.
NeuroNCAP Score[9] ↑	UniAD	**2.030**	1.310	**2.180**	2.430	1.310	1.210	1.110	1.130	1.460
	VAD	2.110	1.160	1.780	2.230	1.130	**1.540**	**1.500**	**1.580**	1.610
	SparseDrive	1.690	1.180	1.820	2.290	0.940	1.140	1.030	0.740	1.250
	DiffusionDrive	1.940	**1.510**	1.960	**2.530**	**1.760**	0.920	1.290	0.950	1.490
Collision Rate(%) ↓	UniAD	**0.722**	**0.888**	**0.640**	**0.565**	0.885	0.954	0.975	0.948	0.863
	VAD	0.777	0.944	0.760	0.652	0.942	**0.886**	**0.887**	**0.871**	0.853
	SparseDrive	0.805	0.944	0.680	0.608	0.942	0.909	0.937	0.974	0.876
	DiffusionDrive	0.750	**0.888**	**0.640**	**0.565**	**0.800**	1.000	0.912	0.974	0.850
Acceleration (%) ↓	UniAD	0.749	1.370	1.134	0.614	1.213	1.018	0.971	1.029	0.997
	VAD	0.719	1.510	**0.533**	0.510	0.892	1.309	1.311	1.426	1.091
	SparseDrive	**0.560**	1.242	**0.533**	**0.469**	0.725	**0.988**	**0.963**	**1.018**	0.841
	DiffusionDrive	0.633	**1.098**	0.649	0.689	**0.690**	1.183	1.151	1.168	0.962
Gen-NCAP Score ↑	UniAD	33.98	11.57	**40.44**	**43.78**	12.63	8.63	56.65	16.33	31.47
	VAD	**40.28**	**18.53**	35.98	43.75	15.64	**18.48**	**68.97**	**32.05**	39.39
	SparseDrive	19.71	5.790	32.18	39.31	5.960	11.92	51.09	11.81	26.01
	DiffusionDrive	25.58	11.37	36.20	43.73	**20.55**	3.170	48.73	11.61	27.49
GNRS ↑	UniAD	**5.37**	8.20	**11.9**	4.26	**6.52**	2.99	2.79	4.50	5.20
	VAD	2.95	4.65	3.32	3.55	2.42	**4.70**	**4.79**	**5.19**	5.04
	SparseDrive	2.34	5.52	3.42	2.54	1.95	4.19	3.84	3.59	3.83
	DiffusionDrive	4.24	**10.8**	5.07	**6.49**	3.26	4.33	4.65	4.27	5.65

6.2 Gen-NCAP Boosts Performance on Corner Case via Providing Open-Loop Data

Since corner cases are relatively rare in real-world driving and collecting such data is both time-consuming and costly, we tackle this challenge by leveraging synthetic data generated by Gen-NCAP. Specifically, we use this data for open-loop training to improve the robustness of autonomous driving (AD) models in handling rare and safety-critical situations during closed-loop simulation. Table 2 presents the performance of the VAD model after fine-tuning on NeuroNCAP scenarios. The results demonstrate that, after fine-tuning with our generated data, the model not only achieves higher overall scores but also significantly reduces the collision rate, indicating an enhanced ability to cope with challenging corner cases. This shows that scenario-focused synthetic data effectively mitigates the scarcity of real-world corner cases.

Table 2: Quantitative comparison between the original VAD and the VAD fine-tuned(VAD*) using Gen-NCAP.

Method	NeuroNcap Score ↑	CollisionRate ↓
VAD	0.81	0.90
VAD*	**1.24**	**0.85**

6.3 Real-to-sim evaluation in open loop

When conducting evaluations in simulation environments, a critical challenge is the Sim2Real gap, the degree to which simulation outcomes deviate from real-world results. To examine this issue, we conduct an **open-loop** evaluation using data generated by our Gen-NCAP. For a fair comparison, the generated data is aligned with the scenarios present in the NeuroNCAP. As presented in Tab. 3, the average displacement error between the original and generated images remains minimal, indicating high fidelity. Moreover, our generated data achieves comparable or lower collision rates than the original.

Table 3: Evaluation of real-to-sim performance on NeuronCap scenarios using typical planning metrics in an open-loop setting.ADE stands for Average Displacement Error.

Model	Modality	ADE @ T(m) ↓			CR @ T (%) ↓		
		1.0s	2.0s	3.0s	1.0s	2.0s	3.0s
UniAd	Original	0.44	0.75	1.16	0.00	0.12	0.21
	NeuroNcap	0.47	0.80	1.24	0.00	0.12	0.24
	Gen-NCAP	0.45	0.78	1.22	0.00	0.09	0.20
VAD	Original	0.43	0.71	1.01	0.00	0.08	0.11
	NeuroNcap	0.44	0.76	1.16	0.00	0.00	0.08
	Gen-NCAP	0.35	0.63	1.03	0.00	0.00	0.00

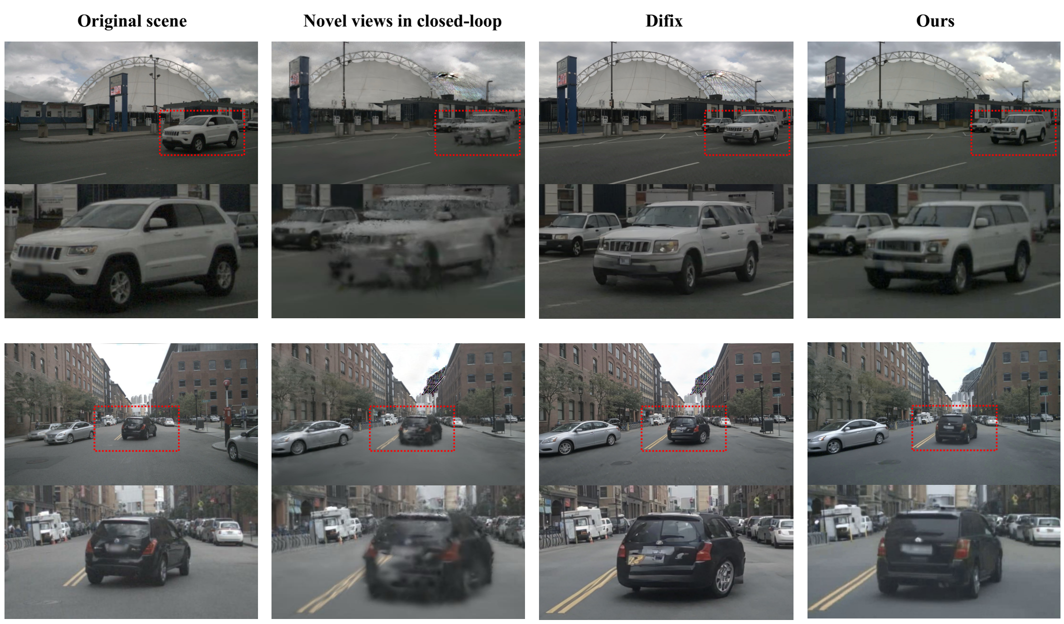
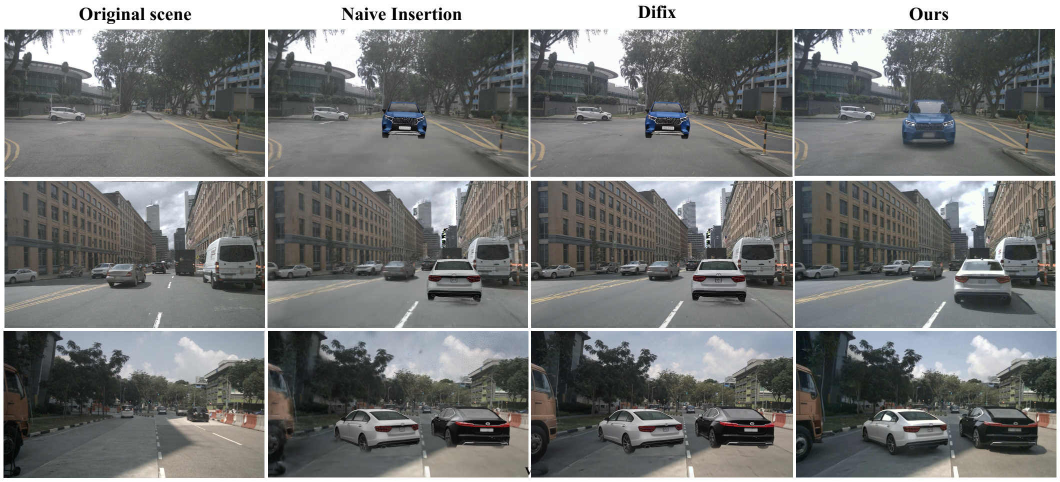


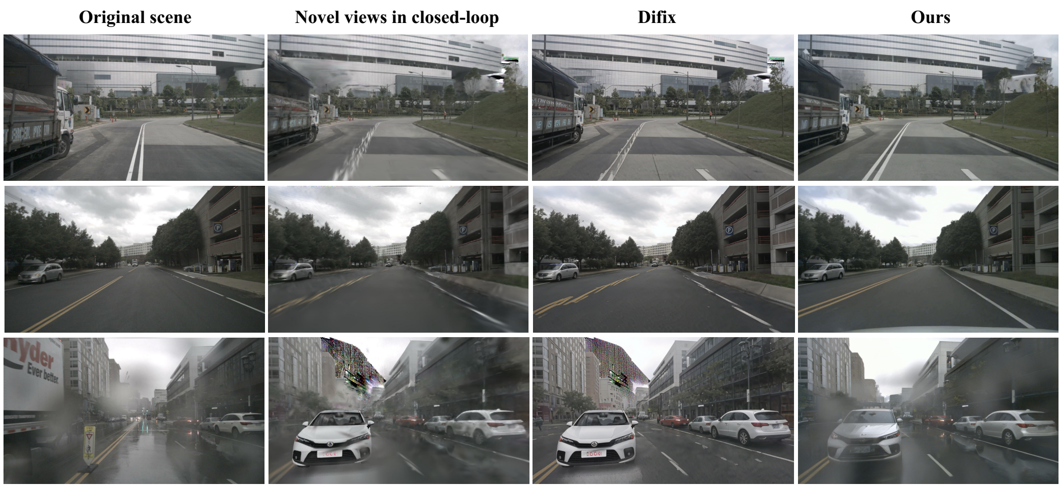
Figure 8: Vehicle motion blur restoration with zoom-in results (below each example).

6.4 Results of the video refinement module

Qualitative results. As shown in Fig. 9, our video refinement module corrects foreground-background misalignment and removes novel-view artifacts. Compared with Difix [31], it achieves better visual quality and remains robust under severe artifacts. As further shown in Fig. 8, Difix mainly improves global image quality but overlooks foreground vehicle restoration.



(a) fix foreground-background misalignment



(b) restore blurriness in novel views

Figure 9: Visualization of the Video Refinement Module's performance in correcting the two types of artifacts. Compared to other methods, our method significantly improves image quality.

Quantitative results. We also provide quantitative results. As shown in Table 4, our method significantly improves the FID compared to images without any post-processing (denoted as the baseline in the table) and clearly outperforms Difix. The FID-A further demonstrates one key advantage of our approach: the ability to restore both background and foreground regions.

Table 4: Quantitative results of video refinement module.

Method	FID ↓	FID-A ↓	FVD ↓
Baseline	77.97	34.46	952.4
Difix	52.63	21.97	1113
Ours	**48.48**	**21.86**	**816.8**

7 Conclusion

In conclusion, Gen-NCAP offers a novel benchmark for comprehensive safety evaluation in autonomous driving by providing a large-scale and diverse corner case dataset alongside a high-fidelity closed-loop simulation engine. Our systematic evaluation reveals significant safety risks in current SoTA E2E-AD models when exposed to challenging and rare corner cases. Additionally, Gen-NCAP's dataset can effectively fine-tune these models, enabling them to substantially improve their robustness, reliability, and overall safety performance in real-world driving scenarios.

References

[1] Yihua Shao, Yeling Xu, Xinwei Long, Siyu Chen, Ziyang Yan, Yang Yang, Haoting Liu, Yan Wang, Hao Tang, and Zhen Lei. Accidentblip: Agent of accident warning based on ma-former. *arXiv preprint arXiv:2404.12149*, 2024.

[2] Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint arXiv:2406.08481*, 2024.

[3] Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*, 2025.

[4] Hao Gao, Shaoyu Chen, Bo Jiang, Bencheng Liao, Yiang Shi, Xiaoyang Guo, Yuechuan Pu, Haoran Yin, Xiangyu Li, Xinbang Zhang, et al. Rad: Training an end-to-end driving policy via large-scale 3dgs-based reinforcement learning. *arXiv preprint arXiv:2502.13144*, 2025.

[5] Chonghao Sima, Kashyap Chitta, Zhiding Yu, Shiyi Lan, Ping Luo, Andreas Geiger, Hongyang Li, and Jose M Alvarez. Centaur: Robust end-to-end autonomous driving with test-time training. *arXiv preprint arXiv:2503.11650*, 2025.

[6] 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.

[7] Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Liu Haiyang, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al. Renderworld: World model with self-supervised 3d label. *arXiv preprint arXiv:2409.11356*, 2024.

[8] Zhihao Guo and Peng Wang. Depth priors in removal neural radiance fields. In *Annual Conference Towards Autonomous Robotic Systems*, pages 367–382. Springer, 2024.

[9] William Ljungbergh, Adam Tonderski, Joakim Johnander, Holger Caesar, Kalle Åström, Michael Felsberg, and Christoffer Petersson. Neuroncap: Photorealistic closed-loop safety testing for autonomous driving. In *European Conference on Computer Vision*, pages 161–177. Springer, 2024.

[10] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *arXiv preprint arXiv:2406.03877*, 2024.

[11] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023.

[12] Euro NCAP. Assessment Protocol – Safety Assist: Collision Avoidance. `https://www.euroncap.com/media/79866/euro-ncap-assessment-protocolsa-collision-avoidance-v104.pdf`, 2023. Version 1.0.4.

[13] Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. *arXiv preprint arXiv:2504.13092*, 2025.

[14] Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. *arXiv preprint arXiv:2504.01941*, 2025.

[15] Fabio Remondino, Ali Karami, Ziyang Yan, Gabriele Mazzacca, Simone Rigon, and Rongjun Qin. A critical analysis of nerf-based 3d reconstruction. *Remote Sensing*, 15(14):3585, 2023.

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

[17] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024.

[18] 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.

[19] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023.

[20] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[21] 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 preprint arXiv:2106.11810*, 2021.

[22] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36:7730–7742, 2023.

[23] Zhihao Guo, Jingxuan Su, Shenglin Wang, Jinlong Fan, Jing Zhang, Liangxiu Han, and Peng Wang. Gp-gs: Gaussian processes for enhanced gaussian splatting. *arXiv preprint arXiv:2502.02283*, 2025.

[24] Junhao Ge, Zuhong Liu, Longteng Fan, Yifan Jiang, Jiaqi Su, Yiming Li, Zhejun Zhang, and Siheng Chen. Unraveling the effects of synthetic data on end-to-end autonomous driving. *arXiv preprint arXiv:2503.18108*, 2025.

[25] Ziyang Yan, Nazanin Padkan, Paweł Trybała, Elisa Mariarosaria Farella, and Fabio Remondino. Learning-based 3d reconstruction methods for non-collaborative surfaces—a metrological evaluation. *Metrology*, 5(2):20, 2025.

[26] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

[27] 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.

[28] Lue Fan, Hao Zhang, Qitai Wang, Hongsheng Li, and Zhaoxiang Zhang. Freesim: Toward free-viewpoint camera simulation in driving scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12004–12014, 2025.

[29] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 822–832, 2025.

[30] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. *arXiv preprint arXiv:2411.19548*, 2024.

[31] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26024–26035, 2025.

[32] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[33] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

[34] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023.

[35] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[36] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al. Omnire: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024.

[37] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[38] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[39] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[40] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[41] Shiwei Liu, Tianlong Chen, Zhenyu Zhang, Xuxi Chen, Tianjin Huang, Ajay Jaiswal, and Zhangyang Wang. Sparsity may cry: Let us fail (current) sparse neural networks together! *arXiv preprint arXiv:2303.02141*, 2023.