

WorldCorner: Evaluating Driving World Models for Safety-Critical Event Generation

Jiawei Wang^{*1}, Bodong Bi^{*1}, Kaiyang Luo¹, Xujie Song¹, Shaowei Hua¹,
Haowei Sun², Jun Gao^{1,3}, Henry X. Liu^{†1} [‡]

¹University of Michigan, ²Laplace Intelligence, ³NVIDIA

Abstract

Driving world models promise scalable synthesis of corner cases that are rare but critical for autonomous vehicle deployment. However, the ability to generate corner cases does not guarantee their quality, and existing evaluations of these models largely focus on normal driving scenarios. We introduce **WorldCorner**, a benchmark for evaluating driving world models on safety-critical event generation. We first build **CornerScenes**, a dataset of more than 1,200 safety-critical scenarios adapted from nuScenes, each paired with its original normal scene. Four complementary dimensions are then designed to evaluate the generated videos: *generation quality*, *controllability*, *critical event realization*, and *ADS functional utility*, complemented by a human study. Benchmarking results indicate broad performance degradation on corner cases and a tendency to fall back on normal-driving priors, and show that conventional quality metrics cannot capture the faithful generation of the intended safety-critical event. WorldCorner establishes a foundation for evaluating corner case generation in world models, and shifts the target from perceptual realism toward safety-oriented fidelity for physical AI systems.

Introduction

Corner cases remain one of the primary obstacles to the large-scale deployment of autonomous vehicles (AVs) (Chen et al. 2024a; Liu and Feng 2024). Such events rarely occur in the real world, making them costly, inefficient, and potentially unsafe to collect at scale. With recent advances in video generation, driving world models promise high-fidelity data synthesis with naturalistic traffic behaviors and photorealistic sensor observations (Wang et al. 2024b; Ren et al. 2025). Particularly, such models can be steered to generate safety-critical corner cases, suggesting a new paradigm for scalable AV safety evaluation (Zhao et al. 2025; Gao et al. 2024).

However, the ability to generate corner cases alone is insufficient. The decisive factor lies in the data quality, and therefore rigorous evaluation of driving world models is essential. General video-generation and world-model evaluations primarily assess visual realism, physical plausibility, and temporal or dynamic consistency (Huang et al. 2025; Li

et al. 2026; Duan et al. 2025), while driving-specific benchmarks additionally consider domain-related criteria, such as adherence to ego-motion conditions and utility for downstream perception and planning (Arai et al. 2024; Zhou et al. 2026; Liang et al. 2026). Yet these evaluations mainly focus on normal driving scenarios, leaving the corner cases most relevant to AV safety largely unexamined. Works that explicitly generate corner cases, by contrast, are mostly limited to evaluating whether a safety-critical event occurs (Xu et al. 2025; Fu et al. 2025). This leaves a key question unanswered: under corner cases, do generation quality and control adherence degrade, is the intended safety-critical event faithfully realized, and can an AV system recognize it?

To answer this question, we introduce **WorldCorner**, a multi-dimensional benchmark for evaluating driving world models on safety-critical scenarios, with four complementary dimensions: 1) *generation quality* measures visual fidelity and consistency; 2) *controllability* measures how well generated results follow scenario layout conditions; 3) *critical event realization* evaluates whether the intended safety-critical event is faithfully reproduced; and 4) *ADS functional utility* evaluates whether an end-to-end autonomous driving system (ADS) can correctly interpret the safety-relevant context. The first two dimensions focus on whether performance degrades under safety-critical conditions, while the latter two examine event-level fidelity and downstream utility. The four dimensions comprise 12 metrics in total, which we report separately rather than aggregate, as they capture distinct and non-monotonically related capabilities.

Such an evaluation requires realistic and controllable corner cases, which existing datasets rarely provide. We therefore construct **CornerScenes** from the real nuScenes dataset (Caesar et al. 2020), focusing on safety-critical interactions between the ego vehicle and other traffic participants, generated with a pretrained behavior simulation model (Zhong et al. 2023). Two types of interactions are considered, depending on whether the ego vehicle or the other participant acts as the adversarial actor, enabling evaluation under both victim and attacker settings. Each corner case is paired with the original normal driving scene sharing the same road structure, background traffic, and environment context, providing a controlled reference for isolating the effect of the safety-critical interaction.

Figure 1 illustrates the overall pipeline: paired normal and

^{*}These authors contributed equally.

[†]Corresponding author.

[‡]This manuscript is under submission. Last updated: July 28, 2026.

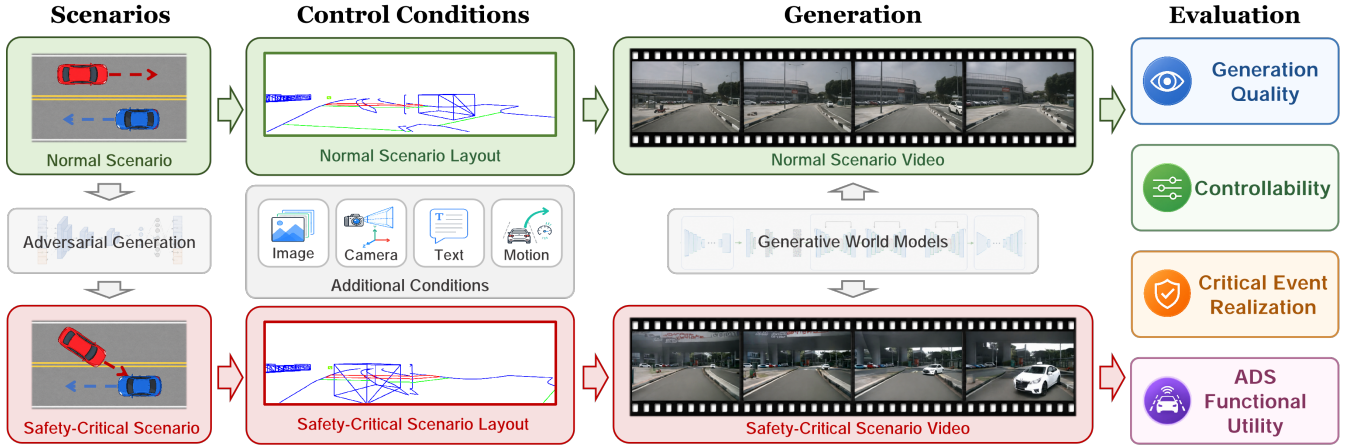


Figure 1: Overview of the **WorldCorner** benchmark pipeline. Behaviors in normal nuScenes scenarios are modified via adversarial generation to construct the safety-critical scenarios in CornerScenes. Control conditions are then extracted, in particular the HD map and the 3D control boxes of the surrounding traffic, which specify the scenario layout and encode the safety-critical interaction. These conditions drive video generation, and the resulting videos are evaluated across four dimensions.

safety-critical scenarios are rendered via generative world models, and the video outputs are assessed across the four dimensions. We evaluate seven representative driving world models, all trained on nuScenes and sharing similar control conditions to ensure a fair comparison. Our main contributions are as follows:

- We introduce **WorldCorner**, the first benchmark dedicated to evaluating driving world models on corner cases with safety-critical interactions.
- We design metrics to systematically measure the corner-case generation ability: generation quality, controllability, critical event realization, and ADS functional utility.
- We build **CornerScenes**, a dataset of more than 1,200 safety-critical scenes adapted from real driving scenes, each paired with its original normal counterpart.
- Benchmarking results reveal broad performance degradation on corner cases and a persistent tendency to fall back on normal-driving priors. Conventional generation metrics imply neither event realization nor functional utility for AV safety evaluation.

Related Works

Video Generation. Video generation models have shown promising capabilities in synthesizing long-horizon, high-resolution, and temporally coherent videos from text or images (Brooks et al. 2024; Yang et al. 2025; Kong et al. 2024). Since they can produce realistic observations of physical environments, such models are increasingly adopted as scalable simulators in physical AI (NVIDIA 2026; Parker-Holder and Fruchter 2025). In autonomous driving, driving world models enable controllable multi-view street-scene generation, conditioned on structured signals such as high-definition maps, 3D bounding boxes, and vehicle trajectories (Gao et al. 2024, 2025; Mei et al. 2024; Guo et al. 2025; Zhao et al. 2025; Chen et al. 2024b; Yang et al. 2026; Ren et al. 2025).

World Model Benchmark. Standardized benchmarks have emerged for systematic evaluation of generative world models. General video benchmarks mostly assess visual realism, semantic alignment, motion quality, and temporal consistency (Huang et al. 2024, 2025; Liu et al. 2024; Wei et al. 2026). World-modeling capabilities are further evaluated for simulating the physical world, including instruction following, physical plausibility, and dynamic consistency (Duan et al. 2025; Li et al. 2026). While these benchmarks provide broad coverage of generation and application metrics, they mostly focus on normal scenarios only. Building on the seminal work of WorldLens (Liang et al. 2026), which focuses on driving world models and assesses generation quality over original normal scenes in the nuScenes dataset (Caesar et al. 2020), we adopt comparable metrics in terms of *generation quality* but measure them under synthesized corner cases and against paired normal references. For controllability, prior work primarily examines adherence to ego-vehicle motion (Arai et al. 2024; Zhou et al. 2026; Liang et al. 2026). This paper, by contrast, evaluates alignment over the entire scene layout and, in particular, over the surrounding participants involved in the safety-critical interaction with the ego vehicle, which captures the corner case generation capability.

Corner Case Generation. The rarity of long-tail scenarios in real-world driving data has driven interest in corner case synthesis via world models (Li et al. 2022a; Song et al. 2023; Xu et al. 2026; Stoler et al. 2024; Zhong et al. 2021). A growing line of work aims at driving video generation with safety-critical interactions, either by adversarial optimization over agent behaviors or scripted hazardous interactions via language models (Xu et al. 2025; Fu et al. 2025; Wang et al. 2025; Zhao et al. 2025; Gao et al. 2024). However, these works typically validate their outputs with a narrow set of proxy metrics, such as collision rate and time-to-collision, which mainly capture whether a safety-critical event happens. By contrast, this paper designs *critical event realization* and

ADS functional utility, which respectively verify whether the intended event is faithfully reproduced and whether the generated data can support AV safety testing.

The WorldCorner Benchmark

To systematically evaluate the ability of driving world models to generate safety-critical corner cases, we introduce WorldCorner. As shown in Figure 2, the benchmark covers four complementary dimensions: *generation quality*, *controllability*, *critical event realization*, and *ADS functional utility*.

The CornerScenes Dataset

Evaluating corner-case generation requires a dataset consisting of realistic safety-critical interactions, which are rare in existing driving datasets. Dedicated accident datasets (Fang et al. 2024) are curated from heterogeneous dashcam and web sources with inconsistent camera configurations, and lack the structured annotations that driving world models require as control conditions, such as HD maps and 3D bounding boxes. Fully synthetic datasets (Wang et al. 2024a) provide such annotations but suffer from a significant sim-to-real gap.

We therefore build CornerScenes by adapting the real nuScenes dataset (Caesar et al. 2020), with only the critical behaviors modified. For each scene, annotators select a vehicle with a potential high-risk interaction with the ego vehicle and assign one of two adversarial roles, depending on whether the ego vehicle or the background vehicle acts as the attacker. Trajectories are then modified with CTG++ (Zhong et al. 2023), a diffusion-based trajectory simulation model trained on nuScenes, and guidance functions are designed to steer toward close contact while preserving behavior plausibility. Results are filtered by rule-based checks and manual review to ensure a valid safety-critical interaction.

The resulting CornerScenes dataset contains over 1,200 scenes, with more than 1,000 for the original training split and 173 for the validation split in nuScenes. Two adversarial roles are roughly evenly distributed. Each corner case is paired with its original normal scene. Since existing world models cannot reliably render actual collisions, every sequence is truncated 3 frames (0.3 s) before the first close-contact frame, so that collision risk is present but no physical contact occurs. The same time window is applied to the paired normal scene.

Note that CornerScenes provides control conditions rather than videos: each scene consists of the modified interaction trajectories together with the original HD map, background traffic, environmental context, and sensor configuration inherited from nuScenes. No ground-truth video exists. Our evaluation is designed accordingly, assessing the generated videos through reference-free visual measures, their alignment with the given control conditions, and their deviation from the intended event. All evaluations in this paper use the validation split. Since the benchmarked models are trained on the nuScenes training split, both the corner cases and their paired normal scenes are unseen at evaluation time, so the paired comparison is free from data leakage. The remaining scenes, adapted from the nuScenes training split, will be released to support future work rather than used for evaluation here. More details are provided in the Supplementary Material.

Generation Quality

This dimension of metrics examines whether world models can synthesize visually authentic, temporally cohesive, and semantically sound driving environments under corner cases. We adopt the generation quality metrics of WorldLens (Liang et al. 2026) and report them on corner cases relative to their paired normal scenes.

G1. Subject Fidelity measures the instance-level visual realism of generated objects in the driving scene. Object regions are cropped from each frame using bounding boxes and scored by class-specific ViT classifiers (Dosovitskiy et al. 2020) pretrained on real data. Higher fidelity indicates stronger category alignment and appearance realism.

G2. Subject Coherence measures the temporal persistence of object identity. Guided by ground-truth track IDs, we extract per-instance feature representations with pretrained re-identification models (He et al. 2021; Zuo et al. 2024) and compute their cross-frame similarity. Higher value indicates better preservation of shape, color, and texture under motion.

G3. Depth Discrepancy measures geometric integrity over time. We estimate monocular depth with Depth Anything V2 (Yang et al. 2024), encode the color-mapped depth maps with a DINOv2 (Oquab et al. 2024) backbone, and report the mean embedding distance between consecutive frames. A lower value indicates more stable scene geometry.

G4. Semantic Consistency measures the temporal stability of scene semantics. We generate per-frame segmentation masks with a pretrained SegFormer (Xie et al. 2021) and compare them across frames in terms of category labeling, regional layout, and class distribution. A higher value indicates more stable semantic structures, with fewer artifacts such as boundary jittering or abrupt label changes.

Controllability

This dimension measures whether a generated video follows its control conditions. Prior work mostly evaluates controllability through ego motion (Arai et al. 2024; Zhou et al. 2026; Liang et al. 2026). By contrast, we evaluate control over the full scene layout, especially the HD map and the 3D control boxes of surrounding traffic, which captures the road structure and interaction behaviors.

C1. Map Layout Alignment checks whether the generated video preserves the road layout in the control condition. At each timestamp, BEVFormer (Li et al. 2022b) combines the six camera views and predicts a BEV map. We compare this prediction with the map in the control condition and report mean Intersection-over-Union (mIoU). A higher score means that the generated video follows the map condition more closely.

C2. 2D Object Alignment evaluates whether controlled traffic participants are rendered at their intended image-plane locations. We apply the YOLO11s (Jocher and Qiu 2024) detector to each frame and match the detections against 3D control boxes projected onto the 2D image. As projected boxes are usually larger than detector outputs, we use Intersection over Detection (IoDet), which measures the fraction of a detected box inside the projected box. A higher score means that the generated video follows the agent conditions more closely at the image level.

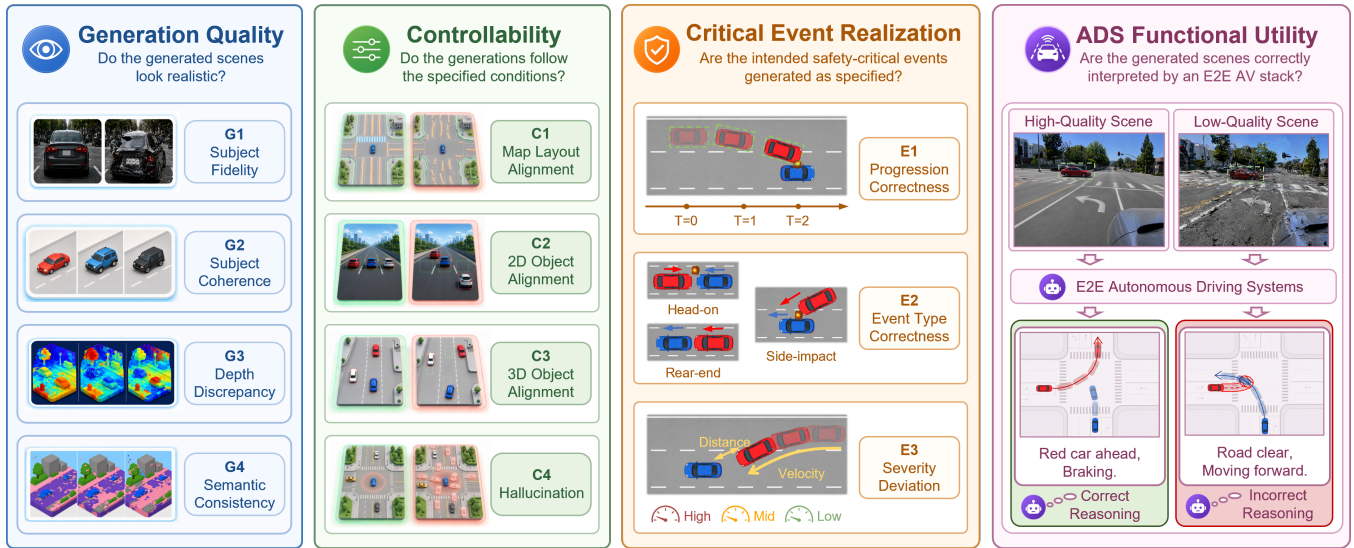


Figure 2: Overview of the **WorldCorner** evaluation framework. Four dimensions are designed to measures corner-case generation capabilities of driving world models: *generation quality*, *controllability*, *critical event realization*, and *ADS functional utility*. These dimensions assess whether generated driving scenes look realistic, follow the specified conditions, reproduce the intended safety-critical events, and can be correctly interpreted by an end-to-end ADS, respectively.

C3. 3D Object Alignment evaluates whether controlled traffic participants are realized as recognizable 3D objects aligned with the specified conditions. We apply a BEV-Former (Li et al. 2022b) model pretrained on nuScenes to each timestamp and match its detections with the 3D boxes in the control condition. Here we report two separate scores: (1) **C3.1 Object Realization Score** treats the specified 3D boxes as ground truth and reports the mean average precision of the detections, and (2) **C3.2 Realized Object Quality Score** measures how closely they follow the specified geometry, motion, and attributes. For both scores, higher values indicate better alignment.

C4. Hallucination counts traffic participants that appear in the generated video but are absent from the control condition. Hallucinated vehicles indicate that scene layout is not fully controllable, and they can alter the outcome of AV safety evaluations by intruding into the intended interaction. Reusing the 3D detections from C3, an object is counted as a hallucination if it matches no box of the same category and lies far from any such box. We report the average number per frame and the number within 10 m of the ego vehicle, both restricted to the drivable area. Lower values are better.

Critical Event Realization

The controllability metrics above test whether the scene layout follows the control conditions frame by frame. Correct realization of a corner case further requires the critical interaction to evolve over time as specified. We therefore evaluate whether the intended event itself is realized, through three metrics that reuse the 3D detections from C3 and focus on a series of key frames near the end of the sequence.

E1. Progression Correctness measures whether the critical vehicle persists throughout the event progression. At each

key frame, the vehicle is counted as recovered if a detection of the same category falls within a BEV center-distance threshold. An event scores 1 only when the adversarial vehicle is recovered for every key frame, and 0 otherwise. We report the mean score over all events, and a higher value indicates that more complete corner-case events are realized.

E2. Event Type Correctness measures whether the generated video renders the same type of safety-critical contact as the control condition. Since the sequences end before actual contact, we extrapolate the trajectories forward and label the first contact as one of three types: *front*, *rear*, or *side*, following the definitions in CTG++ (Zhong et al. 2023). A scene is correct when the labels from the specified trajectories and from the trajectories recovered in the generated video agree. We report the fraction of correct scenes, where a higher rate indicates closer adherence to the intended event type.

E3. Severity Deviation measures how far the severity of the generated event deviates from the specified one. Severity is characterized at each key frame by the inter-vehicle distance and the closing speed derived from it. We combine the relative errors of the two quantities into a per-frame deviation score and average it over key frames and scenes. This metric is computed only for scenes in which the adversarial vehicle is completely recovered, as defined in E1. A lower score indicates closer adherence to the specified severity.

ADS Functional Utility

This dimension assesses whether generated corner cases can be correctly interpreted by AVs, which is the ultimate requirement for using such data in AV safety evaluation. If the safety-relevant content is faithfully represented, an ADS should recognize the conflict and plan an evasive response. What we measure is interpretability rather than driving com-

petence: the planned trajectory need not physically avoid the collision, since an attempted evasion already shows that the scene conveys the risk and thus constitutes a valid corner case. Conversely, a planner that proceeds as if nothing were happening reveals that the generated scene has failed to deliver the safety-critical information.

Specifically, we design a score of **U1. Risk Unawareness**, which sums the overlap areas between the bird’s-eye-view footprints of the ego and adversarial vehicles across all waypoints of the ADS planned trajectory. We deploy three state-of-the-art open-source end-to-end ADS systems, UniAD, SparseDrive, and DiffusionDrive (Hu et al. 2023; Sun et al. 2025; Liao et al. 2025), and average across them so that the score reflects the generated scene rather than the capability of an individual ADS system. A lower value indicates that the ADS recognizes the safety-relevant content and produces a more appropriate response.

Human Evaluation

All metrics above are rule-based, and can be automatically calculated by machines. Human judgment is instead composite, weighing multiple factors at once. For *ADS functional utility*, for instance, our rule reduces the planned trajectory to a footprint overlap. By contrast, a human can observe the overall behavior, such as braking, yield, or steering away, and can therefore tell whether the generated scene was understood by the ADS system. Such considerations arise in every dimension, and no fixed rule captures all of them. We therefore complement the automated protocol with a human study that lets this integrative judgment be expressed directly. This also allows us to check whether the rule-based scores and model rankings agree with human annotation.

The study mirrors the same four dimensions. The evaluation interface can overlay the input conditions (3D boxes and map layout) onto the generated video and replay the ADS planned trajectory in BEV. Annotators answer four mandatory questions then ask annotators to holistically rate the sample across the four dimensions on a 0–5 scale. The full guideline and evaluation results are provided in the Supplementary Material.

Experiments

We evaluate seven representative driving world models based on the WorldCorner framework, including MagicDrive (Gao et al. 2024), MagicDrive-v2 (Gao et al. 2025), DriveDreamer-2 (Zhao et al. 2025), DreamForge (Mei et al. 2024), DiST-4D (Guo et al. 2025), OpenDWM (Chen et al. 2024b), and X-Scene (Yang et al. 2026). These models cover the main approaches to driving video generation with structured control conditions. Due to space constraints, more details are provided in the Supplementary Material.

Per-Aspect Evaluations

Generation Quality. Figure 3 compares each model on CornerScenes against its paired normal driving scenes. All four metrics degrade consistently but moderately, indicating that safety-critical interactions make video generation more challenging while basic visual quality is largely preserved. The

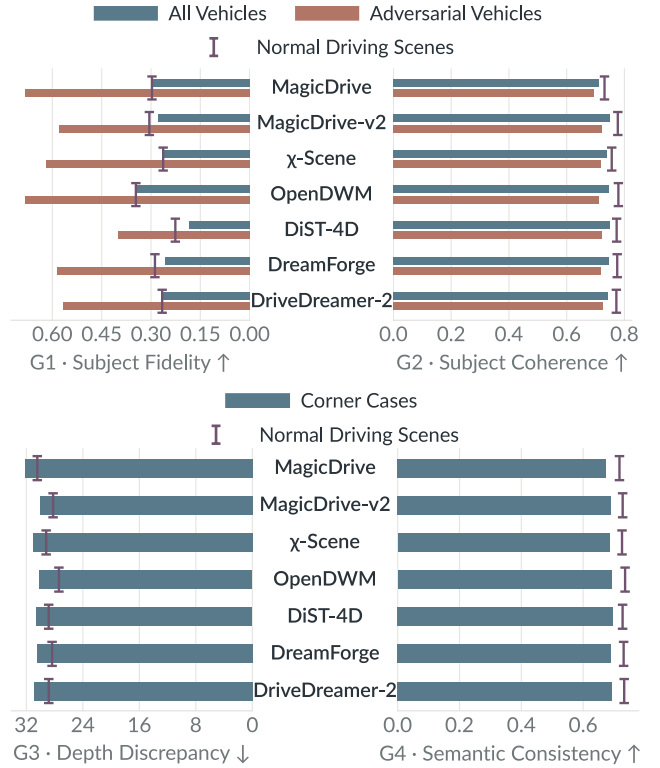


Figure 3: Benchmarking results for **Generation Quality**.

relative ordering of models remains roughly unchanged, suggesting that stronger generation quality in normal driving scenes tends to carry over to corner cases. Differences across models lie mainly in subject fidelity. OpenDWM performs best, likely benefiting from its large-scale multi-dataset training. Although DiST-4D, DriveDreamer-2, and DreamForge are all conditioned on real initial frames, DiST-4D appears more sensitive to inconsistencies between the initial frames in nuScenes and the control conditions in CornerScenes, causing its low rank in fidelity. We also report subject-level results separately for all vehicles and for the adversarial vehicle alone. Fidelity for the adversarial vehicle is much higher, likely because it stays close to the ego vehicle and occupies a large, well-resolved image region.

Controllability. Figure 4 shows that the degradation from normal scenes to corner cases is far more significant in controllability than in generation quality. Map layout alignment (C1) and object alignment over all vehicles, in both 2D (C2) and 3D (C3), drop substantially, indicating that reproducing the road structure and the motion relations among participants is considerably harder once the interaction deviates from normal driving patterns. The drop is most severe for C3.2 on the adversarial vehicle: even when the critical vehicle is realized, its geometry, position, and motion deviate markedly from the specified condition, which is precisely the information that determines whether the safety-critical event occurs as intended. Model rankings also shift considerably between the two settings, so strong controllability on

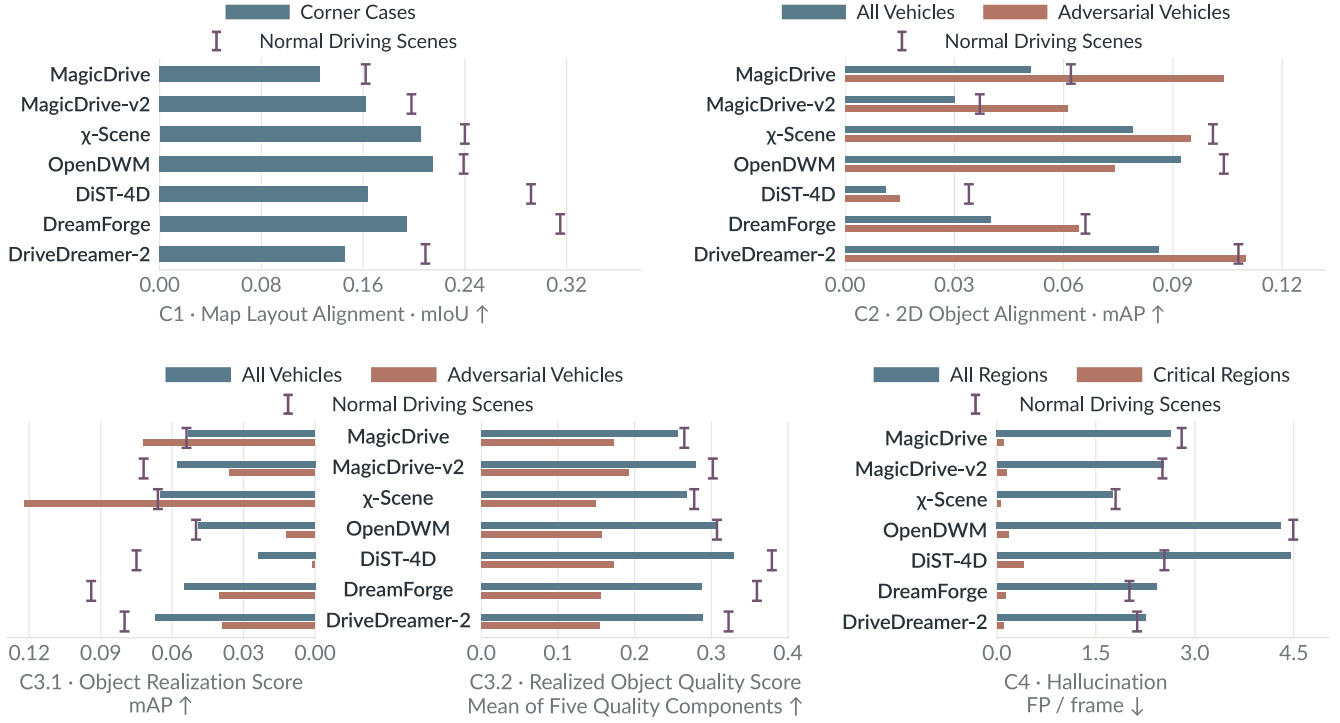


Figure 4: Benchmarking results for **Controllability**.

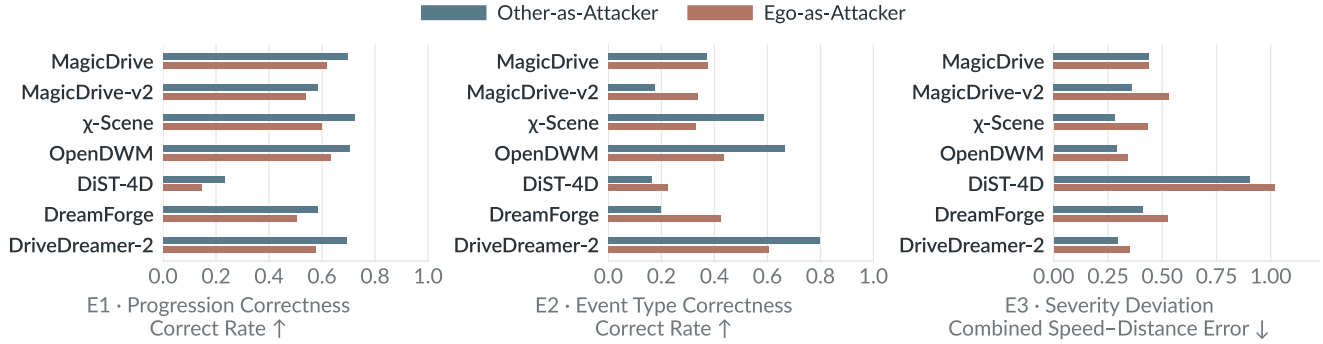


Figure 5: Benchmarking results for **Critical Event Realization**.

normal scenes does not carry over to corner cases. Hallucination (C4), by contrast, remains comparable between normal and corner cases. Most models generate roughly one to four hallucinated vehicles per frame, suggesting that hallucination is a general limitation of current models rather than a corner-case-specific failure.

Critical Event Realization. Figure 5 shows that current models still struggle to realize safety-critical events: the mean progression correctness rate is roughly 54%–66% for most models, and four of the seven fall below 40% on event-type correctness. X-Scene, OpenDWM, and DriveDreamer-2 perform best overall among the three metrics. Comparing the two setups of the attacker role, most models achieve better event-level performance when another vehicle acts as the at-

tacker, likely because the ego trajectory then remains closer to the training distribution. This dimension shows that object recovery and control adherence captured in the previous two dimensions do not guarantee faithful realization of the entire event. Existing structured controls can constrain scene elements and spatial layouts per frame, but faithfully realizing a safety-critical event requires explicitly modeling how relative motion evolves over time.

ADS Functional Utility. Figure 6 shows that overlap areas under corner cases significantly exceed those of the paired normal scenes, confirming that CornerScenes poses a substantial challenge for existing AV systems. What this dimension primarily reveals, however, is not planning skill but scene interpretability: for the same scene generated by dif-

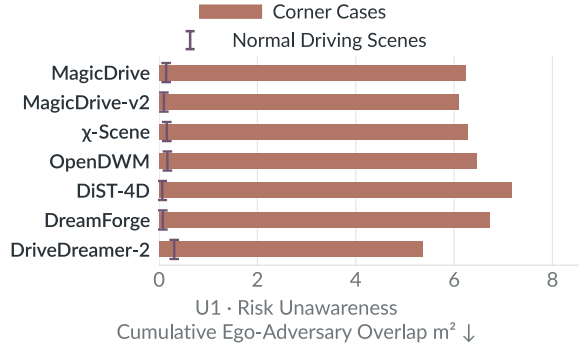


Figure 6: Benchmarking results for **ADS Functional Utility**.



(a) The adversarial vehicle is recognized and avoided.



(b) The adversarial vehicle is not recognized.

Figure 7: Representative cases of the **ADS Functional Utility** dimension. Each example shows the generated multi-view scene alongside the BEV visualization. The blue and purple vehicles represent the ego and the adversarial vehicle, respectively. The original trajectory is shown in green and the ADS-planned trajectory is shown in red.

ferent driving world models, a smaller overlap means the ADS has understood the risk and attempted an evasive maneuver. Along this direction, DriveDreamer-2 produces the scenes whose risk is most reliably conveyed to the ADS. Figure 7 illustrates two representative examples. In Figure 7(a), the adversarial vehicle on the right is rendered with visible artifacts yet remains recognizable, and the ego vehicle steers away for collision avoidance. In Figure 7(b), severe distortion prevents the ADS from recognizing a valid traffic participant ahead at all, and the planned trajectory proceeds forward with no avoidance response. The contrast shows that a generated scene need not be visually perfect to be functionally usable, but the safety-critical participant must survive the end-to-end ADS pipeline to be acted upon.

Observations & Insights

Normal-scene performance masks corner-case failures.

Across most metrics, the evaluated world models consistently perform worse on corner cases than on paired normal

driving scenes. This gap is attributable to the interaction itself rather than to scene composition, which remains fixed during scenario modification. Evaluating on normal scenes alone therefore overstates a model’s readiness for reliable synthesis of safety-critical scenarios. Dedicated corner-case evaluation is necessary.

Object realization does not imply event realization. High visual quality and successful realization of the critical vehicle do not translate into faithful reproduction of event type and severity. Since a high-risk interaction is defined by how traffic behaviors evolve over time, the correct presence and appearance of controlled objects are necessary but not sufficient for event fidelity. Both generation models and evaluation frameworks must explicitly capture multi-agent interaction.

Corner-case generation is bottlenecked by learned scene priors. Video generation models follow control conditions poorly when those conditions require deviating from normal driving patterns, with event realization degrading most when the ego vehicle acts as the attacker. This suggests that current models continue the motion priors learned from training data rather than realize the intended abnormal interactions. Enabling robust synthesis of safety-critical events calls for broader coverage of corner cases in training data and stronger structured-control mechanisms.

Perceptual realism does not imply functional utility. Model rankings on the first two dimensions have no monotonic relationship to ADS functional utility. A usable scene need not reproduce every visual detail. This reframes a central question for synthetic data: *how real is real enough?* The answer depends not on absolute perceptual realism but on whether the physical AI system recognizes the critical information. For safety evaluation, such information must survive generation and remain recoverable throughout the downstream decision pipeline. Driving world models should therefore optimize for decision-relevant fidelity at safety-critical events rather than perceptual realism alone.

Conclusion

This paper introduces WorldCorner, a standardized benchmark for evaluating driving world models on corner case generation capabilities across four complementary dimensions. Our results reveal a clear gap between what conventional metrics measure on normal scenes and what dedicated corner-case evaluation exposes. Strong generation quality neither guarantees faithful realization of the intended safety-critical event nor ensures that the generated data can be interpreted by an AV system. Together with the CornerScenes dataset and the accompanying human study, WorldCorner provides a foundation for benchmarking world model capabilities toward reliable safety validation of physical AI systems.

References

- Arai, H.; Ishihara, K.; Takahashi, T.; and Yamaguchi, Y. 2024. Act-bench: Towards action controllable world models for autonomous driving. *arXiv preprint arXiv:2412.05337*.
- Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.;

- Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.
- Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.
- Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2024a. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10164–10183.
- Chen, R.; Wu, Z.; Liu, Y.; Guo, Y.; Ni, J.; Xia, H.; and Xia, S. 2024b. UniMLVG: Unified Framework for Multi-view Long Video Generation with Comprehensive Control Capabilities for Autonomous Driving. *arXiv preprint arXiv:2412.04842*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Duan, H.; Yu, H.-X.; Chen, S.; Fei-Fei, L.; and Wu, J. 2025. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 27713–27724.
- Fang, J.; Li, L.-l.; Zhou, J.; Xiao, J.; Yu, H.; Lv, C.; Xue, J.; and Chua, T.-S. 2024. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22030–22040.
- Fu, Y.; Zha, R.; Tian, P.; and Di, X. 2025. Llm-based realistic safety-critical driving video generation. *arXiv preprint arXiv:2507.01264*.
- Gao, R.; Chen, K.; Xiao, B.; Hong, L.; Li, Z.; and Xu, Q. 2025. MagicDrive-V2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 28135–28144.
- Gao, R.; Chen, K.; Xie, E.; Hong, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2024. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*, volume 2024, 22841–22860.
- Guo, J.; Ding, Y.; Chen, X.; Chen, S.; Li, B.; Zou, Y.; Lyu, X.; Tan, F.; Qi, X.; Li, Z.; et al. 2025. Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 27231–27241.
- He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15013–15022.
- Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17853–17862.
- Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.
- Huang, Z.; Zhang, F.; Xu, X.; He, Y.; Yu, J.; Dong, Z.; Ma, Q.; Chanpaisit, N.; Si, C.; Jiang, Y.; et al. 2025. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jocher, G.; and Qiu, J. 2024. Ultralytics YOLO11.
- Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuan-video: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.
- Li, D.; Fang, Y.; Chen, Y.; Yang, S.; Cao, S.; Wong, J.; Luo, M.; Wang, X.; Yin, H.; Gonzalez, J.; et al. 2026. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38.
- Li, K.; Chen, K.; Wang, H.; Hong, L.; Ye, C.; Han, J.; Chen, Y.; Zhang, W.; Xu, C.; Yeung, D.-Y.; et al. 2022a. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision*, 406–423. Springer.
- Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; and Dai, J. 2022b. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*.
- Liang, A.; Kong, L.; Yan, T.; Liu, H.; Yang, Y.; Huang, Z.; Yin, W.; Zuo, J.; Hu, Y.; Zhu, D.; et al. 2026. WorldLens: Full-spectrum evaluations of driving world models in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 36385–36399.
- Liao, B.; Chen, S.; Yin, H.; Jiang, B.; Wang, C.; Yan, S.; Zhang, X.; Li, X.; Zhang, Y.; Zhang, Q.; et al. 2025. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12037–12047.
- Liu, H. X.; and Feng, S. 2024. Curse of rarity for autonomous vehicles. *nature communications*, 15(1): 4808.
- Liu, Y.; Cun, X.; Liu, X.; Wang, X.; Zhang, Y.; Chen, H.; Liu, Y.; Zeng, T.; Chan, R.; and Shan, Y. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22139–22149.
- Mei, J.; Hu, T.; Yang, X.; Wen, L.; Yang, Y.; Wei, T.; Ma, Y.; Dou, M.; Shi, B.; and Liu, Y. 2024. Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003*.
- NVIDIA. 2026. Cosmos 3: Omnimodal World Models for Physical AI. *arXiv preprint arXiv:2606.02800*.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2024. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*.

- Parker-Holder, J.; and Fruchter, S. 2025. Genie 3: A new frontier for world models. URL <https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/>. Blog post.
- Ren, X.; Lu, Y.; Cao, T.; Gao, R.; Huang, S.; Sabour, A.; Shen, T.; Pfaff, T.; Wu, J. Z.; Chen, R.; et al. 2025. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*.
- Song, Q.; Tan, K.; Runeson, P.; and Persson, S. 2023. Critical scenario identification for realistic testing of autonomous driving systems. *Software Quality Journal*, 31(2): 441–469.
- Stoler, B.; Navarro, I.; Jana, M.; Hwang, S.; Francis, J.; and Oh, J. 2024. Safeshift: Safety-informed distribution shifts for robust trajectory prediction in autonomous driving. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, 1179–1186. IEEE.
- Sun, W.; Lin, X.; Shi, Y.; Zhang, C.; Wu, H.; and Zheng, S. 2025. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 8795–8801. IEEE.
- Wang, J.; Sun, H.; Yan, X.; Feng, S.; Gao, J.; and Liu, H. X. 2025. TeraSim-World: Worldwide Safety-Critical Data Synthesis for End-to-End Autonomous Driving. *arXiv preprint arXiv:2509.13164*.
- Wang, T.; Kim, S.; Ji, W.; Xie, E.; Ge, C.; Chen, J.; Li, Z.; and Luo, P. 2024a. DeepAccident: A Motion and Accident Prediction Benchmark for V2X Autonomous Driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5599–5606.
- Wang, X.; Zhu, Z.; Huang, G.; Chen, X.; Zhu, J.; and Lu, J. 2024b. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, 55–72. Springer.
- Wei, J.; He, Y.; Hou, Y.; Qi, H.; Wang, S.; Shi, J.; Li, K. F.; Zheng, Z.; and Wu, W. 2026. Ref4D-VideoBench: Four-Dimensional Reference-Based Evaluation of Text-to-Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7719–7729.
- Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.
- Xu, R.; Lin, H.; Jeon, W.; Feng, H.; Zou, Y.; Sun, L.; Gorman, J.; Tolstaya, K.; Tang, S.; White, B.; et al. 2026. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3709–3718.
- Xu, Z.; Li, B.; Gao, H.-a.; Gao, M.; Chen, Y.; Liu, M.; Yan, C.; Zhao, H.; Feng, S.; and Zhao, H. 2025. Challenger: Affordable adversarial driving video generation. *arXiv preprint arXiv:2505.15880*.
- Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.
- Yang, Y.; Liang, A.; Mei, J.; Ma, Y.; Liu, Y.; and Lee, G. H. 2026. X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. *Advances in Neural Information Processing Systems*, 38: 104415–104451.
- Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2025. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, 83048–83077.
- Zhao, G.; Wang, X.; Zhu, Z.; Chen, X.; Huang, G.; Bao, X.; and Wang, X. 2025. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10412–10420.
- Zhong, Z.; Rempe, D.; Chen, Y.; Ivanovic, B.; Cao, Y.; Xu, D.; Pavone, M.; and Ray, B. 2023. Language-guided traffic simulation via scene-level diffusion. In *Conference on robot learning*, 144–177. PMLR.
- Zhong, Z.; Tang, Y.; Zhou, Y.; Neves, V. d. O.; Liu, Y.; and Ray, B. 2021. A survey on scenario-based testing for automated driving systems in high-fidelity simulation. *arXiv preprint arXiv:2112.00964*.
- Zhou, Y.; Shao, H.; Wang, L.; Zong, Z.; Li, H.; and Waslander, S. L. 2026. DrivingGen: A Comprehensive Benchmark for Generative Video World Models in Autonomous Driving. *arXiv preprint arXiv:2601.01528*.
- Zuo, J.; Nie, Y.; Zhou, H.; Zhang, H.; Wang, H.; Guo, T.; Sang, N.; and Gao, C. 2024. Cross-video identity correlating for person re-identification pre-training. *Advances in Neural Information Processing Systems*, 37: 25228–25250.