

Human Behavior-Informed Crash Scenario Generation with Real-World Crash Priors for Autonomous Vehicle Safety Evaluation

Mingxing Peng¹, Xusen Guo¹, Long Chen², Xintao Yan³, Siyu Teng⁴, and Jun Ma^{1,*}

¹The Hong Kong University of Science and Technology (Guangzhou), China

²Institute of Automation Chinese Academy of Sciences, Beijing, China

³The University of Hong Kong, Hong Kong, China

⁴Shenzhen University, Shenzhen, China

*corresponding author: jun.ma@ust.hk

ABSTRACT

Reliable safety evaluation of autonomous vehicles (AVs) is essential to improving road safety, yet it depends critically on realistic simulation of rare crashes. Existing crash scenario generation methods can increase collision occurrence, but often fail to realistically reproduce how crashes evolve before impact or the distribution of crash types observed in the real world. Here, we present CrashSim, a human behavior-informed crash scenario generation framework that uses real-world crash priors to guide generative multi-agent traffic simulation for more reliable AV safety evaluation. These priors capture how real-world crashes evolve before impact and how different crash types are distributed, allowing limited crash data to guide realistic and scalable scenario generation across naturalistic driving contexts. We evaluate CrashSim against competing methods, showing that it more closely reproduces real-world pre-impact behavior, collision dynamics, collision geometry and crash-type distributions. We further use CrashSim to construct nuCrash dataset, containing over 4,000 crash and near-crash scenarios. Closed-loop evaluation of five AV planners shows that nuCrash more effectively exposes differences in planner safety capabilities than nuScenes. An LLM-assisted evaluation agent further analyzes planner failures to provide capability-level diagnoses and targeted improvement guidance. Together, CrashSim enables realistic and scalable crash generation for more informative AV safety evaluation.

Keywords: Crash Scenario Generation; AV Safety Evaluation; Generative AI; Large Language Models; Diffusion Models

Introduction

Road traffic safety remains a fundamental societal challenge. As autonomous driving rapidly develops and becomes increasingly integrated into modern transportation systems, reliable safety evaluation of autonomous vehicles is essential not only for improving road safety, but also for supporting the broader deployment and continued development of autonomous driving technologies¹⁻³. However, safety-critical events, particularly crashes and severe near-crashes, are extremely rare in real-world driving. Reliable assessment of vehicle safety requires sufficient data on such events, but they account for only a very small proportion of real-world driving data^{4,5}, even in datasets comprising millions of trips or tens of millions of miles of driving, as illustrated in Fig. 1a. Consequently, adequately assessing vehicle safety through real-world road testing alone can require a very large amount of driving, making the evaluation process both time-consuming and costly⁶. To enable more efficient evaluation of rare and hazardous interactions, safety-critical scenario generation in simulation has become an important component of autonomous vehicle safety evaluation⁷, and maintaining realism in both driving behavior and scene-level distributions is critical to reliable safety evaluation.

Existing studies have developed diverse approaches for exposing autonomous vehicles to safety-critical interactions in simulation. Environment-based testing promotes the occurrence of rare and safety-critical events through targeted scenario design or accelerated sampling⁸⁻¹³; optimization-based methods optimize scenario parameters or agent behaviours towards more critical or collision-inducing interactions¹⁴⁻¹⁶; generative approaches combine learned naturalistic traffic distributions with guidance towards safety-critical outcomes¹⁷⁻²⁰; and recent LLM-based methods directly exploit the reasoning capabilities of LLMs to generate collision trajectories²¹⁻²³. Together, these advances have substantially improved the efficiency of safety-critical testing. However, existing methods commonly prioritize increasing collision occurrence or scenario criticality, while giving limited attention to modelling the interaction process preceding impact. This limitation is particularly important because real-world crashes typically emerge through evolving interactions rather than as isolated collision outcomes. As a result, existing

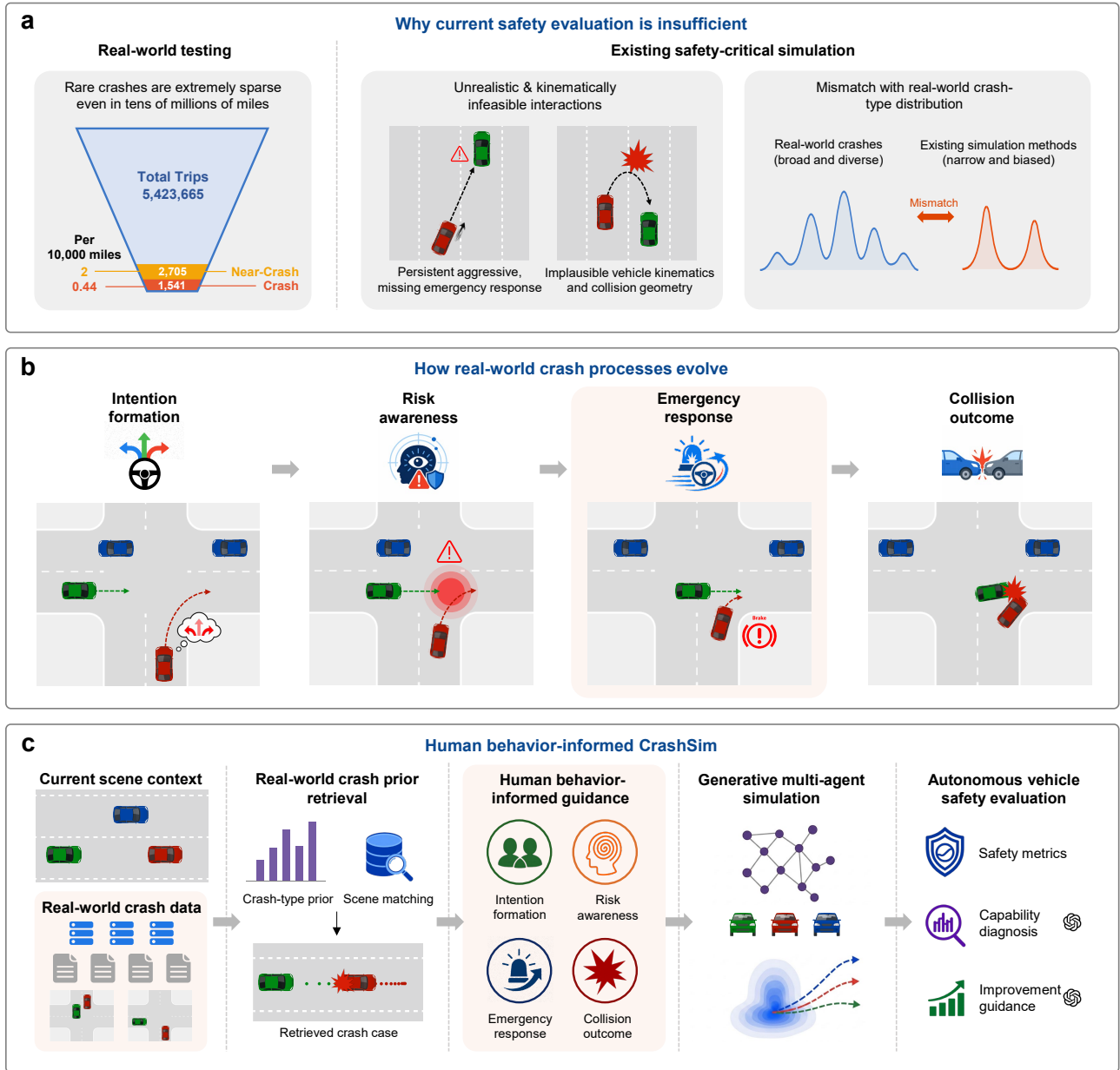


Figure 1. Challenges in autonomous vehicle safety evaluation and overview of CrashSim. **a** Limitations of current autonomous vehicle safety evaluation, including the sparsity of real-world crashes and the behavioral, kinematic, and distributional biases of existing safety-critical simulation. Statistics in the real-world testing funnel are derived from ref.⁵. **b** Real-world crash processes evolve from intention formation through risk awareness and emergency response to the eventual collision outcome whereas existing safety-critical simulation often overlooks the emergency response and leads to aggressive collisions. **c** CrashSim retrieves real-world crash priors based on the current scene context and uses them to provide human behavior-informed guidance for generative multi-agent simulation and autonomous vehicle safety evaluation.

methods may generate crashes characterized by persistent aggressive motion, missing emergency responses or implausible vehicle kinematics. Moreover, existing approaches rarely explicitly account for the scene-level distribution of generated crashes, which can cause generated scenario sets to over-represent particular conflict types and deviate from the composition of real-world crashes. As illustrated in Fig. 1a, these limitations highlight two dimensions of realism in crash scenario generation: the pre-collision interaction process within individual scenarios and the crash-type distribution across the generated scenario set.

Real-world crash data provide a basis for addressing the two dimensions of realism discussed above. At the level of a single

crash scenario, crashes typically develop through a temporally evolving process rather than occurring as isolated terminal events. Analyses of real crashes and near-crashes indicate that drivers exhibit diverse braking and evasive responses before a collision^{24,25}. Models of human collision avoidance further characterize this process as a dynamic interplay between threat detection, response selection, and evasive action execution²⁶. To characterize this temporal evolution for crash simulation, we represent crash development through four stages: intention formation, risk awareness, emergency response, and collision outcome, as illustrated in Fig. 1b. At the level of the scenario set, real crashes also exhibit a structured distribution across interaction types. Based on real-world crash databases, the National Highway Traffic Safety Administration (NHTSA) developed a pre-crash typology that organizes collision-producing interactions according to their preceding manoeuvres and conflict configurations²⁷. The resulting statistics reveal a highly non-uniform distribution across crash types, with longitudinal conflicts, lane-change and lateral interactions, intersection encounters, and opposing-direction conflicts occurring at markedly different frequencies. Real-world crash data thus serve as a valuable source of information for guiding the construction of realistic crash simulations.

However, learning such real-world crash characteristics directly from crash data remains challenging. In contrast to naturalistic driving datasets, which provide abundant and relatively standardized trajectories for training generative traffic models, real-world crash data are sparse, unevenly distributed across interaction types, and heterogeneous in representation. Such heterogeneous crash data span structured crash reports and taxonomies²⁷, naturalistic crash and near-crash records²⁸, reconstructed trajectories²⁹, and crash video recordings^{30,31}. Only a limited fraction of these data provide sufficiently complete and consistent trajectories for direct generative modelling, making it difficult to train a general-purpose crash generator that captures the diversity of real-world collision processes and crash types. Retrieval-augmented generation (RAG) offers a practical strategy for leveraging such sparse and heterogeneous crash data by retrieving contextually relevant information at inference time, without requiring all available crash information to be represented in a homogeneous training corpus³². Meanwhile, large language models (LLMs) and vision-language models (VLMs) have demonstrated strong capabilities for reasoning over complex scene context and heterogeneous forms of prior knowledge^{33,34}, and have increasingly been applied to safety-critical driving scenario generation^{21–23}. Inspired by these developments, we leverage retrieval-based crash priors together with the semantic reasoning capabilities of LLMs and VLMs to incorporate heterogeneous real-world crash data into generative modelling without relying on large-scale, uniformly structured crash trajectories.

Here we present CrashSim, a novel framework that incorporates real-world crash priors into generative multi-agent traffic simulation (Fig. 1c). Given a naturalistic driving scene, the VLM agent interprets the road context and traffic interactions among nearby road users, together with real-world crash-type statistics, to identify a potential safety-critical interaction. CrashSim then retrieves a relevant real-world crash prior through hierarchical matching of interaction type, event context, and temporal motion pattern. The retrieved crash prior is transformed into human behavior-informed guidance that preserves its relative-motion pattern and behavior-stage timing while adapting them to the geometry and interaction context of the current scene. This guidance steers the diffusion-based generator, pretrained on naturalistic driving data, towards realistic safety-critical interactions. Because real-world crash data are inherently sparse, CrashSim organizes the available cases into a small-scale crash retrieval library rather than relying on a large crash trajectory dataset for model training. By transferring the crash priors contained in this library to diverse naturalistic driving scenes, CrashSim can generate a substantially larger set of crash scenarios while preserving both realistic temporal evolution at the individual-scenario level and the real-world distribution of crash types across the generated scenario set.

We show that CrashSim improves realism at both the individual-scenario and scenario-set levels, more closely reproducing real-world pre-impact behavior and collision characteristics while also matching the observed distribution of crash types. We further use CrashSim to construct nuCrash, a dataset of more than 4,000 crash and near-crash scenarios that enriches rare safety-critical interactions while retaining these behavioral and distributional characteristics. Closed-loop evaluation of five autonomous-driving planners shows that nuCrash more clearly differentiates their safety performance than naturalistic nuScenes. An LLM-assisted evaluation agent further relates these differences to planner-specific capability limitations and failure patterns and provides targeted improvement guidance, thereby supporting interpretable safety evaluation. In contrast to existing autonomous-driving evaluation frameworks that primarily rely on predefined quantitative metrics^{35,36}, our approach provides a capability-oriented interpretation of planner performance under safety-critical interactions. Together, these results demonstrate that sparse real-world crash priors can be transferred across naturalistic driving contexts to improve the realism of safety-critical scenario generation, while LLM-assisted analysis enables more informative and interpretable autonomous-vehicle safety evaluation.

Results

This study models real-world crashes as temporally evolving processes comprising four stages: intention formation, risk awareness, emergency response and collision outcome, rather than treating collision occurrence as a single generation objective. We first show that CrashSim more faithfully reproduces the temporal progression, pre-impact dynamics, response timing

and collision behaviors observed in real-world crashes, while also better preserving real-world crash-type distributions than competing scenario-generation methods. We then use CrashSim to construct nuCrash, a long-tail safety-critical dataset that substantially increases the prevalence and diversity of crash and near-crash scenarios while retaining key behavioral and distributional characteristics of real-world crashes. Closed-loop evaluation of five autonomous driving planners shows that nuCrash reveals substantially stronger planner-specific performance differences than naturalistic nuScenes, and that the LLM-based evaluation agent can translate deterministic metrics and scenario-level evidence into capability assessments, failure diagnoses and actionable improvement guidance. Finally, ablation studies show that real-world crash prior retrieval primarily improves the fidelity of collision-process dynamics, and VLM-based scene understanding improves the alignment of generated crash-type distributions with real-world data.

Experimental setup

Dataset. CrashSim generates realistic crash scenarios from naturalistic driving contexts under the guidance of real-world crash priors. We use nuScenes³⁷ as the source of naturalistic initial scenes. The dataset contains 1,000 urban driving scenes of 20 s each, from which we extract the motion trajectories of the ego vehicle and surrounding traffic participants together with the associated road context. To characterize how real-world crashes evolve and how different crash types are distributed, we incorporate two sources of crash prior knowledge^{27,29}. We use reconstructed crash trajectories²⁹, derived from the SHRP2 NDS^{28,38}, to characterize the pre-collision evolution and vehicle dynamics of real-world crashes. The NHTSA pre-crash scenario typology²⁷ provides complementary empirical statistics on the distribution of real-world crash types. Using these naturalistic driving contexts and real-world crash priors, we construct the nuCrash dataset with CrashSim. nuCrash comprises over 4,000 safety-critical scenarios, including crashes and near-crashes across diverse longitudinal and lateral interactions. Its crash-category distribution broadly follows that observed in real-world crash data and spans diverse long-tail collision scenarios, enabling safety evaluation across a wide range of crash types. We further compare nuCrash with SimEngine³⁹, an existing long-tail driving dataset.

Scenario-generation baselines. We compare CrashSim with two safety-critical scenario-generation methods. Strive¹⁵ uses a graph-based variational autoencoder to learn a traffic-motion prior and generates collision scenarios through latent-space adversarial optimization, whereas LD-scene¹⁹ uses LLM-generated guidance to steer a diffusion model towards adversarial interactions. Both methods primarily optimize towards collision outcomes rather than modeling how real-world crashes evolve through a temporal collision process, which can result in overly aggressive and behaviorally implausible collisions.

Evaluation metrics. We evaluate collision scenario realism from three aspects: collision-process fidelity, behavioral plausibility and scene-level distributional realism. Collision-process fidelity is assessed using the temporal composition of intention formation, risk awareness, emergency response and collision, together with the temporal evolution of risk awareness and emergency response, pre-impact longitudinal acceleration, reaction time, response time and braking severity. Behavioral plausibility is evaluated from collision impact locations and the proportions of adversarial and other non-ego vehicles leaving the drivable area. Scene-level distributional realism is assessed across sixteen fine-grained crash types and five high-level conflict categories, with Total Variation (TV) distance⁴⁰ and Jensen-Shannon divergence (JSD)⁴¹ quantifying discrepancies from real-world distributions. For dataset-level comparison, we additionally report crash and near-crash rates to characterize the prevalence of safety-critical events.

Planner evaluation. We evaluate five autonomous driving planners spanning rule-based and learning-based approaches: Lane-graph planner⁴², PDM-closed planner⁴³, imitation learning-based (IL) planner⁴⁴, imitation learning-based multi-agent motion (IL-Multi) planner⁴⁵ and Diffusion planner⁴⁶. Each planner is evaluated in closed loop on both naturalistic nuScenes scenarios and safety-critical nuCrash scenarios. Following the temporal setting of the nuScenes motion planning task, at each replanning step, the planner uses 2 seconds (4 steps) of motion history to predict a 6 seconds (12 steps) future trajectory, with replanning performed every 0.5 s based on the updated scene state. An LLM-based evaluation agent further analyzes deterministic metrics and scenario-level evidence across the two datasets to assess planner capabilities, diagnose failure mechanisms, and identify directions for planner improvement.

CrashSim reproduces realistic human behavior-informed collision processes

Real-world collision scenarios cannot be characterized by a single driving action, but instead unfold through a temporally evolving sequence of intention formation, risk awareness and emergency response. We therefore investigate whether the generated scenarios reproduce this complete behavioral progression together with the associated driving dynamics and collision outcomes as shown in Fig. 2.

CrashSim more faithfully reproduces the temporal progression observed in real-world crashes than competing methods. Fig. 2a, b shows that CrashSim preserves the temporal allocation of intention formation, risk awareness, emergency response and collision, while recovering the gradual transition from risk perception to emergency action before impact. In contrast,

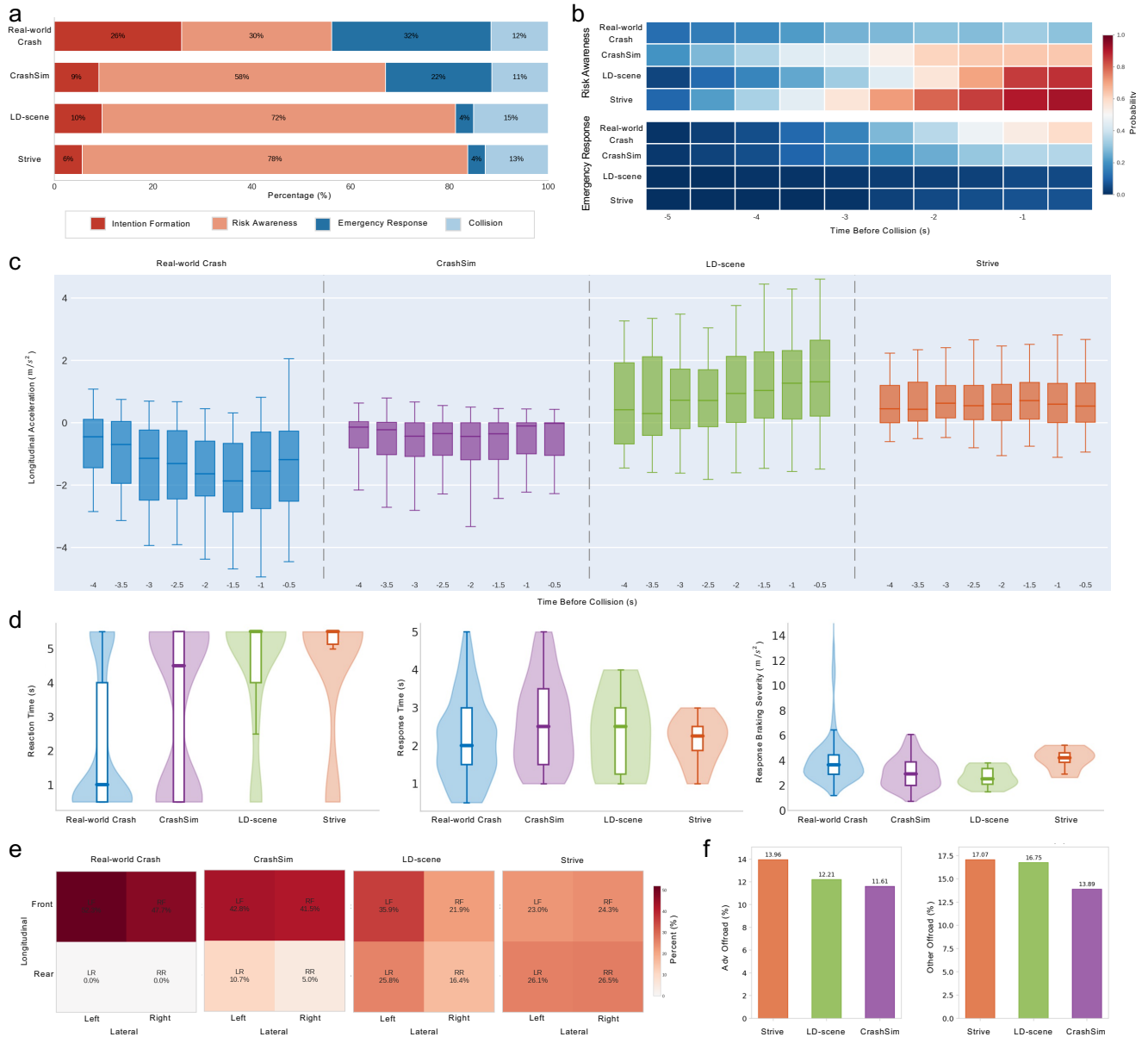


Figure 2. CrashSim generates realistic human behavior-informed collision processes. **a** Temporal composition of intention formation, risk awareness, emergency response and collision in real-world crashes and generated scenarios. CrashSim retains the complete four-stage progression and more closely reproduces the stage proportions observed in real-world crashes. **b** Time-resolved probability heatmaps showing the evolution of risk awareness and emergency response before collision. CrashSim captures the gradual development of risk awareness followed by the emergence of emergency response. **c** Temporal distributions of longitudinal acceleration during the final four seconds preceding collision. CrashSim reproduces the predominantly decelerative response and its temporal evolution in real-world crashes. **d** Distributions of reaction time response and braking severity across real-world crashes and generated scenarios. Together **a–d** show that CrashSim captures key human behavioral dynamics underlying crash progression including risk development response timing and emergency braking. **e** Spatial distributions of collision impact locations across the front, rear, left and right regions of the ego vehicle. CrashSim generates behaviorally realistic ego-at-fault collisions, characterized by a predominance of frontal impacts, thereby enabling more effective evaluation of ego-planner safety. **f** Percentages of adversarial and other non-ego vehicles that leave the drivable area. The low off-road rates produced by CrashSim indicate physically plausible pre-collision interactions and stronger compliance with road boundaries. Together, **e–f** demonstrate the behavioral plausibility of CrashSim-generated collision scenarios.

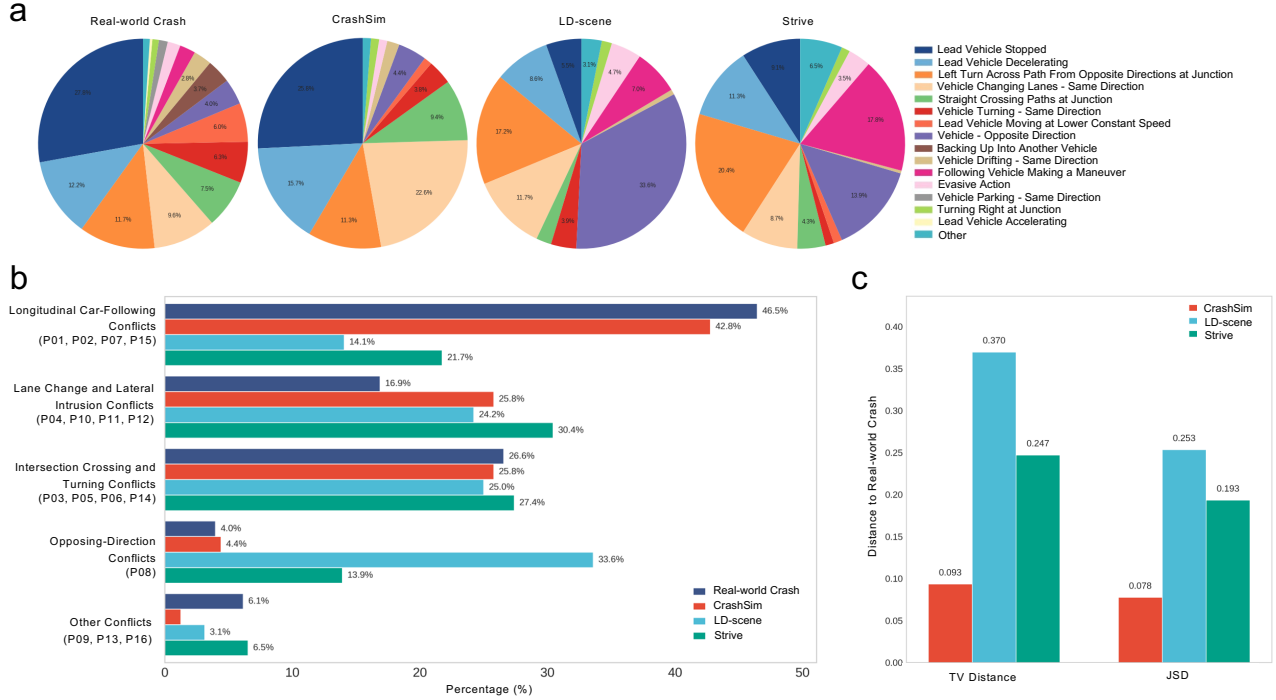


Figure 3. CrashSim preserves realistic crash scenario distributions. **a** Distribution of fine-grained crash scenarios across real-world crashes and generated datasets. **b** Distribution of five high-level crash categories, including longitudinal car-following conflicts, lane-change and lateral intrusion conflicts, intersection crossing and turning conflicts, opposing-direction conflicts, and other conflicts. **c** Distributional discrepancy from real-world crashes, quantified using total variation distance for the five high-level categories and Jensen-Shannon divergence for the sixteen fine-grained crash types. Across both levels of semantic granularity, CrashSim most closely reproduces real-world crash statistics and achieves the lowest distributional discrepancy among the compared methods.

LD-scene and Strive remain in the risk-awareness stage for prolonged periods and exhibit little emergency response immediately before collision, resulting in overly aggressive and behaviorally implausible collision processes. This difference is also reflected in the longitudinal dynamics shown in Fig. 2c. Real-world crashes exhibit increasingly negative longitudinal acceleration as the collision approaches, and CrashSim reproduces this decelerative tendency, whereas LD-scene and Strive maintain predominantly positive acceleration into the conflict with the ego vehicle. Fig. 2d further compares the distributions of reaction time and response time, together with braking severity during the emergency-response stage. Across these measures, CrashSim shows closer overall agreement with the real-world distributions in both response timing and braking intensity. Together, these results indicate that CrashSim reproduces the evolving behavioral and dynamic processes preceding a collision, rather than merely producing the terminal collision outcome.

CrashSim also generates collision behaviors that more closely resemble those observed in real-world crashes. Fig. 2e shows that the impact-location distribution generated by CrashSim is highly consistent with real-world crash statistics, particularly in reproducing the predominance of frontal impacts characteristic of ego-at-fault collisions, thereby enabling more effective evaluation of ego-planner safety. Fig. 2f further shows that adversarial and surrounding vehicles are substantially less likely to contact the curb before collision than in scenarios generated by competing methods. Together, these results demonstrate that CrashSim improves both behavioural realism and kinematic feasibility, yielding more effective safety-critical driving scenarios for evaluating autonomous driving planners.

Collectively these results show that CrashSim reproduces realistic collision processes through temporally coherent human responses realistic driving dynamics and plausible collision outcomes. This enables the generation of safety-critical scenarios with substantially higher behavioral realism than existing scenario generation methods.

CrashSim preserves realistic crash scenario distributions

Generating realistic individual crashes does not necessarily imply that the resulting scenario set reflects the diversity of real-world traffic conflicts. We therefore evaluate whether CrashSim reproduces the distribution of crash scenarios observed in naturalistic crash datasets across both fine-grained crash types and higher-level conflict categories, as summarized in Fig. 3.

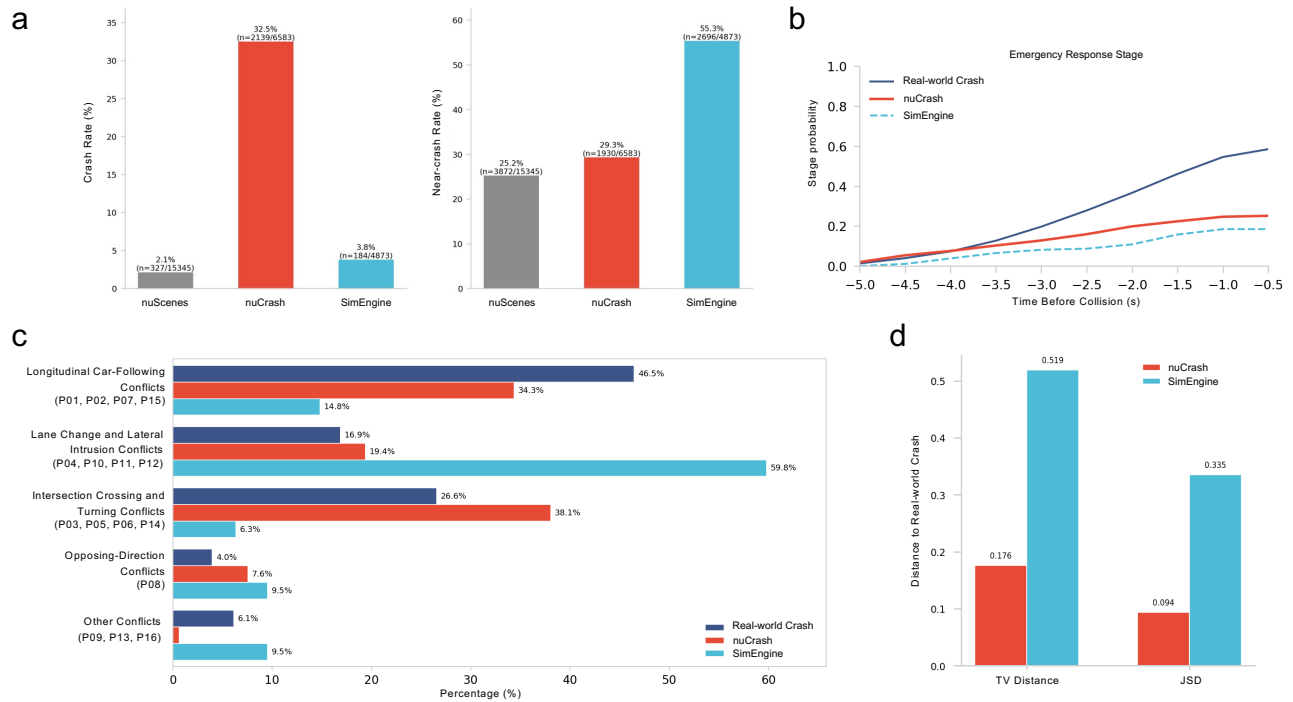


Figure 4. nuCrash constructs a realistic long-tail crash dataset. **a** Comparison of crash rate and near-crash rate among nuScenes, nuCrash, and SimEngine. **b** Temporal evolution of the emergency response stage before collision in real-world crashes, nuCrash, and SimEngine. **c** Distribution of high-level crash categories in real-world crashes, nuCrash, and SimEngine. **d** Distributional discrepancy from real-world crashes, measured using Total Variation distance and Jensen-Shannon divergence. **a** shows that nuCrash substantially increases the prevalence of realized collisions relative to the naturalistic nuScenes dataset, whereas SimEngine is dominated by near-crash cases and contains relatively few actual collisions. **b-d** demonstrate that nuCrash more faithfully reproduces the temporal dynamics and scenario distributions observed in real-world crashes than the existing long-tail dataset. Collectively, these results show that nuCrash provides a more realistic and challenging dataset for autonomous vehicle safety evaluation.

CrashSim closely reproduces the distribution of fine-grained crash scenarios observed in real-world data. As shown in Fig. 3a, the proportions of sixteen representative crash types generated by CrashSim align well with those in the real-world dataset, whereas LD-scene and Strive deviate substantially across multiple scenario categories. The relatively high proportion of Vehicle Changing Lanes scenarios reflects the scene-conditioned generation process of CrashSim. Although the real-world crash-type distribution is incorporated as an auxiliary prior during scenario reasoning, the identification of potential safety-critical interactions remains primarily grounded in the current scene context. Since the original nuScenes scenes contain a relatively large number of interactions between vehicles in adjacent lanes, lane-change-related conflicts are correspondingly more frequent in the generated scenarios. At the aggregated level, Fig. 3b shows that CrashSim most faithfully preserves the relative frequencies of longitudinal car-following, lane-change, intersection, opposing-direction and other conflicts. Together, these results demonstrate that CrashSim can generate safety-critical scenarios across diverse driving contexts while remaining aligned with real-world crash-type distributions.

We further quantify distributional realism using TV distance and JSD between generated and real-world crash distributions. Fig. 3c shows that CrashSim consistently achieves the smallest distributional discrepancy under both metrics, substantially outperforming LD-scene and Strive. The improvement under two distributional measures demonstrates that CrashSim more faithfully preserves the overall statistical composition of real-world crash scenarios.

Collectively, these results demonstrate that CrashSim not only generates realistic individual collision scenarios but also faithfully reconstructs the population-level distribution of real-world crashes. This capability enables more representative, diverse and statistically realistic safety-critical scenario generation for autonomous vehicle evaluation.

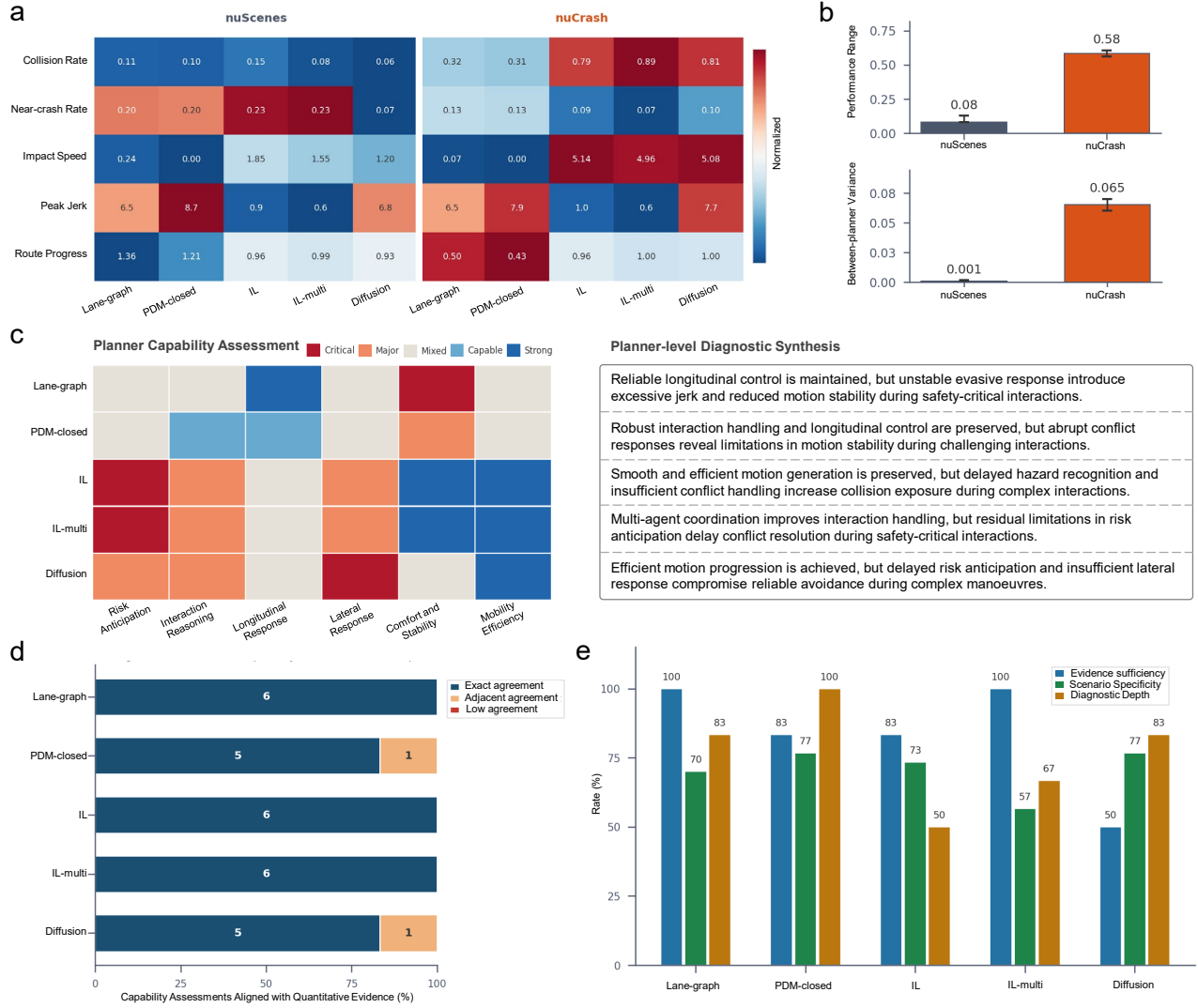


Figure 5. Interpretable and reliable LLM-based evaluation of autonomous driving planners. **a** Planner performance across five safety and driving-quality metrics on nuScenes and nuCrash. Values are normalized within each metric for visual comparison. **b** Performance range and between-planner variance on the two datasets, showing substantially stronger planner differentiation on nuCrash. **c** LLM-based capability assessment across six planner capabilities together with planner-level diagnostic synthesis that translates quantitative evaluation evidence into interpretable planner strengths and limitations. **d** Alignment between capability assessments and quantitative evidence. **e** Planner-level audit results quantifying the evidence sufficiency, scenario specificity, and diagnostic depth. Together these analyses demonstrate that the LLM-based evaluation agent can provide interpretable and reliable planner assessments.

nuCrash provides a realistic long-tail crash dataset

Naturalistic driving datasets, such as nuScenes, faithfully capture everyday traffic interactions but contain relatively few safety-critical events, limiting their coverage of the long-tail conditions encountered in real-world autonomous driving. Using scenes from nuScenes as initial scenarios and CrashSim as the scenario generation framework, we construct nuCrash, a large-scale safety-critical dataset comprising diverse crash and near-crash scenarios. As shown in Fig. 4, we compare nuCrash with both the original nuScenes dataset and SimEngine, an existing synthetic long-tail dataset.

Compared with both nuScenes and SimEngine, nuCrash contains a substantially larger proportion of safety-critical scenarios, making it a more challenging long-tail crash dataset. As shown in Fig. 4a, nuCrash achieves the highest crash rate among the three datasets while maintaining a balanced composition of crash and near-crash events. In contrast, nuScenes contains only a limited number of hazardous interactions, whereas SimEngine is dominated by near-crash scenarios.

Beyond increasing scenario criticality, nuCrash also preserves the behavioral and statistical characteristics of real-world

a Learning-based: IL Planner	b Rule-based: Lane-graph Planner
Cross-dataset diagnostic synthesis On nuScenes, IL retains strong interaction reasoning and longitudinal response. On nuCrash, risk anticipation is weaker, amplifying a visible weakness in nuScenes. Behavioral degradation centers on delayed or absent responses, with secondary inadequate lateral avoidance.	Cross-dataset diagnostic synthesis On nuScenes, Lane-graph retains strong interaction reasoning and longitudinal response. On nuCrash, interaction reasoning and lateral response degrade in a long-tail-specific pattern. Behavioral degradation centers on unstable evasive manoeuvres.
Guidance rationale Observed limitation Limitation amplified on long-tail nuCrash: risk anticipation failures are concentrated, particularly at intersections, where delayed or absent responses lead to frequent collisions. Possible mechanism The planner may struggle to predict and prioritize imminent risks in complex, high-speed environments, which could contribute to delayed braking or absent evasive actions.	Guidance rationale Observed limitation Long-tail-specific degradation on nuCrash: comfort and stability collapse during high-speed evasive manoeuvres with excessive jerk and lateral instability. Failures are concentrated, particularly at intersections. Possible mechanism The planner may prioritize collision avoidance over stability, which could contribute to abrupt and poorly controlled evasive actions. This could result from inadequate trajectory smoothing or insufficient lateral control tuning.
Planner refinement Improvement guidance Add stronger response-time supervision and additional lead-vehicle deceleration training coverage for lead-vehicle deceleration conflicts. Performance-limiting scenarios Performance is most limited in dense, medium-speed longitudinal interactions involving lead-vehicle deceleration and delayed or absent responses; a second cluster appears under high-speed, dense interactions at intersections.	Planner refinement Improvement guidance Add trajectory smoothing and stability-aware control constraints for evasive manoeuvres in lead-vehicle deceleration conflicts. Performance-limiting scenarios Performance is most limited in dense, low-speed longitudinal interactions involving lead-vehicle deceleration and unstable evasive manoeuvres; a second cluster appears under dense, low-speed intersection conflicts associated with unstable evasive manoeuvres.

Figure 6. Qualitative case studies of LLM-based planner evaluation. **a** Learning-based IL Planner. **b** Rule-based Lane-graph Planner. The LLM-based evaluation agent synthesizes cross-dataset planner behavior across naturalistic and long-tail scenarios. It identifies performance degradation and possible failure mechanisms. It further translates the diagnostic findings into planner refinement guidance and highlights representative performance-limiting scenarios. These cases illustrate the ability of the LLM-based evaluation to provide interpretable and actionable planner assessments across distinct planner paradigms.

crashes. As shown in Fig. 4b, the temporal evolution of emergency responses in nuCrash closely tracks that observed in the real-world crash dataset, whereas SimEngine exhibits markedly different response dynamics. Fig. 4c further shows that nuCrash reproduces the distribution of high-level crash categories across longitudinal, lane-change, intersection, opposing-direction, and other conflicts. In contrast, SimEngine substantially overrepresents lane-change conflicts while underrepresenting several other categories. These observations are quantitatively confirmed by the TV distance and JSD to the real-world crash distribution shown in Fig. 4d. nuCrash achieves consistently smaller distances than SimEngine under both metrics, demonstrating that it enriches safety-critical events without compromising the behavioral and distributional realism of real-world crashes.

Collectively, these results demonstrate that nuCrash combines the realism of naturalistic crash data with substantially richer safety-critical interactions. By preserving both the behavioral dynamics and statistical characteristics of real-world crashes while significantly increasing the prevalence of hazardous scenarios, nuCrash provides a realistic and challenging long-tail crash dataset for autonomous vehicle safety evaluation.

CrashSim enables interpretable autonomous vehicle safety evaluation

Beyond generating safety-critical scenarios, CrashSim supports interpretable safety evaluation by combining closed-loop planner performance with LLM-based diagnostic analysis. Fig. 5a shows clear differences in planner behavior between naturalistic and long-tail scenarios. These differences become substantially more pronounced on nuCrash. Fig. 5b quantifies this effect. The normalized performance range increases from 0.08 on nuScenes to 0.58 on nuCrash while the between-planner variance rises from 0.001 to 0.065. The increased separation reflects distinct planner-specific weaknesses under safety-critical interactions rather than uniform degradation. Learning-based planners mainly exhibit increased collision severity while rule-based planners show more pronounced degradation in motion stability or route completion.

Fig. 5c shows that the LLM-based evaluation agent translates this quantitative evidence into capability assessments and planner-level diagnostic summaries. For example, as shown in Fig. 5c, the IL planner is assessed as having strong comfort and stability but critical limitations in risk anticipation. This assessment is consistent with the quantitative results in Fig. 5a where IL exhibits low peak jerk but substantially increased crash rate on nuCrash. The diagnostic summaries further associate these

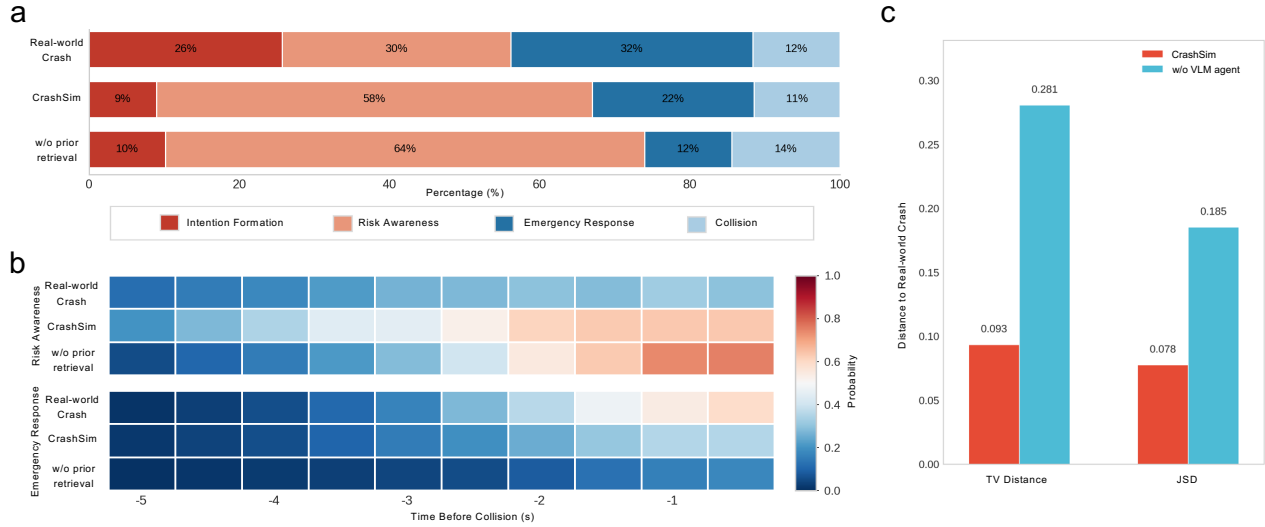


Figure 7. Ablation studies of the crash-prior retrieval module and the VLM agent. **a** Temporal composition of the collision process in real-world crashes, CrashSim and CrashSim without the crash-prior retrieval module. Removing the crash-prior retrieval module increases the discrepancy from the real-world composition, indicating that retrieved crash knowledge helps preserve realistic collision-process characteristics. **b** Temporal evolution of risk awareness and emergency response probabilities for the same settings. The complete framework more closely reproduces the temporal progression observed in real-world crashes, demonstrating that the crash-prior retrieval module improves the realism of pre-collision behavioral dynamics. **c** Total variation and Jensen-Shannon divergence between generated and real-world crash distributions for CrashSim and CrashSim without the VLM-based scene-understanding agent. The complete framework achieves consistently lower distances, indicating that incorporating real-world crash-type distribution priors into scene-level reasoning improves the alignment of generated scenario types with real-world crash statistics.

capability limitations with concrete behavioral patterns such as delayed hazard recognition or unstable evasive responses. The resulting assessments therefore characterize not only whether planner performance deteriorates but also which capabilities are most affected under long-tail conditions.

The reliability of these assessments is examined in Fig. 5d and Fig. 5e. Fig. 5d shows that the capability assessments are highly consistent with the corresponding quantitative evidence across planners. The planner-level audit in Fig. 5e further shows generally high evidence sufficiency and diagnostic depth while revealing some variation in scenario specificity across planners. These results indicate that the LLM-based evaluation remains closely grounded in quantitative evidence while providing planner-specific and interpretable diagnostic insights.

Fig. 6 presents case studies of the LLM-based planner evaluation for representative learning-based and rule-based planners. For IL planner, the dominant degradation on nuCrash is weakened risk anticipation with delayed or absent responses emerging as a major failure pattern particularly in rear-end collision scenarios triggered by lead-vehicle deceleration. The evaluation agent further attributes this vulnerability to insufficient exposure of the learning-based planner to long-tail collision patterns during training and consequently recommends stronger response-time supervision and additional training coverage for lead-vehicle deceleration conflicts. The Lane-graph planner exhibits a different limitation with long-tail degradation dominated by unstable evasive manoeuvres and excessive jerk. The corresponding guidance emphasizes trajectory smoothing and stability-aware control constraints. The identified performance-limiting scenarios further enable targeted validation of whether the proposed improvement guidance effectively mitigates the diagnosed planner weaknesses. Together, these cases demonstrate that CrashSim extends safety evaluation beyond performance measurement by connecting long-tail failures with capability limitations and actionable directions for planner refinement.

Ablation studies

We next examined the respective contributions of the real-world crash-prior retrieval module and the VLM-based scene-understanding agent. The crash-prior retrieval module retrieves behavioral and kinematic evidence from real crash cases. The VLM agent interprets the current driving context and identifies the most plausible adversarial vehicle and its potential behavior.

As shown in Fig. 7a, removing the crash-prior retrieval module reduced the proportion of emergency response from 22% to 12%, increasing its discrepancy from the real-world value of 32%. Figure 7b further shows that, without crash-prior retrieval,

risk awareness developed later and emergency response remained consistently weak. Incorporating retrieved crash evidence promoted earlier risk recognition and a progressively stronger emergency response as the collision approached. These results indicate that the crash-prior retrieval module enables CrashSim to reproduce collision processes that more closely resemble those observed in real-world crashes.

The VLM agent primarily improved the distributional realism across crash scenario types. Fig. 7c shows that removing this agent increased the TV distance from 0.093 to 0.281 and the JSD from 0.078 to 0.185. The VLM-based agent combines prompts informed by real-world prior distributions with its visual understanding capabilities to analyze the driving scene and identify potential collision patterns consistent with the current driving context. This scene-aware identification improves the agreement between the generated and real-world distributions of crash scenario types.

These ablations reveal the functions of the two components. The crash-prior retrieval module improves the realism of collision-process dynamics, whereas the VLM agent enhances the distributional alignment of crash scenario types. Together, they enable CrashSim to generate crashes that more closely resemble real-world crashes in both collision-process characteristics and scenario-type distributions.

Discussion

CrashSim addresses a central mismatch between safety-critical traffic simulation and the way crashes develop in real-world driving. Existing generators are effective at increasing the frequency of safety-critical outcomes, yet the final outcome does not ensure that the preceding interaction follows a behaviorally plausible progression. By grounding scenario generation in retrieved real-world crash priors, CrashSim shifts the focus from inducing collisions to reproducing behaviorally plausible pre-collision interaction dynamics. The observed agreement in behavioral-stage composition, pre-impact longitudinal dynamics, response timing, impact location and scene-level crash-type distributions indicates that this grounding improves the realism of both individual interactions and the scene-level distribution of crash types.

CrashSim also produces two resources with broader utility: a retrieval database constructed from real-world crash data, which can provide empirical crash priors for other diffusion-based or LLM-based scenario generation methods, and the nuCrash dataset, which extends naturalistic driving evaluation to safety-critical conditions that are rarely represented in conventional driving datasets. nuScenes provides rich contextual observations of routine traffic, whereas nuCrash systematically enriches safety-critical interactions while preserving realistic behavioral characteristics and introducing diverse crash scenarios whose composition reflects real-world crash distributions. The substantially larger between-planner variation observed on nuCrash suggests that it provides a more discriminative setting for safety evaluation than naturalistic driving data alone. In particular, performance differences that remain weakly expressed during predominantly routine driving become more distinguishable when planners are exposed to rapidly evolving hazardous interactions that demand timely and coordinated responses. This increased sensitivity to planner-specific weaknesses highlights the value of nuCrash for evaluating safety performance under challenging interactions that are underrepresented in naturalistic driving data.

CrashSim further enables a capability-oriented interpretation of closed-loop planner evaluation. Rather than treating the LLM-based evaluation agent as an alternative to quantitative assessment, we use it to jointly analyze deterministic trajectory metrics, scenario-level statistics, failure patterns and their interaction contexts. This analysis translates heterogeneous measurements into structured assessments of planner capabilities and supports cross-dataset reasoning about how these capabilities change from naturalistic to safety-critical driving. Such a formulation connects low-level performance measurements with higher-level behavioral interpretation, helping to identify which aspects of planner behavior are most strongly associated with the observed performance degradation and to derive corresponding directions for planner improvement. More broadly, this illustrates how language-model-based reasoning can serve as an analytical layer over structured simulation evidence, while leaving the underlying performance measurements grounded in deterministic evaluation.

A remaining limitation lies in the diagnostic depth of the LLM-based evaluation. Although the current analysis can associate performance changes with observed behavioral patterns and scenario conditions, its attribution of the underlying failure mechanisms remains relatively coarse. Future work should proceed in three directions. First, the LLM-based evaluation should be extended with deeper attribution analysis to better identify the mechanisms underlying planner failures and their relationships with specific scenario conditions. Second, CrashSim can be further strengthened by incorporating broader real-world crash databases and more reactive modelling of surrounding traffic. Third, safety-critical datasets such as nuCrash could be explored as training resources for improving planner robustness to rare events, with particular attention to integrating safety-critical scenarios into planner training and adaptation without degrading performance under naturalistic driving conditions.

Methods

CrashSim is a human behavior-informed framework for generating realistic safety-critical driving scenarios and evaluating autonomous driving planners under long-tail datasets. The framework combines scene understanding, retrieval of real-world

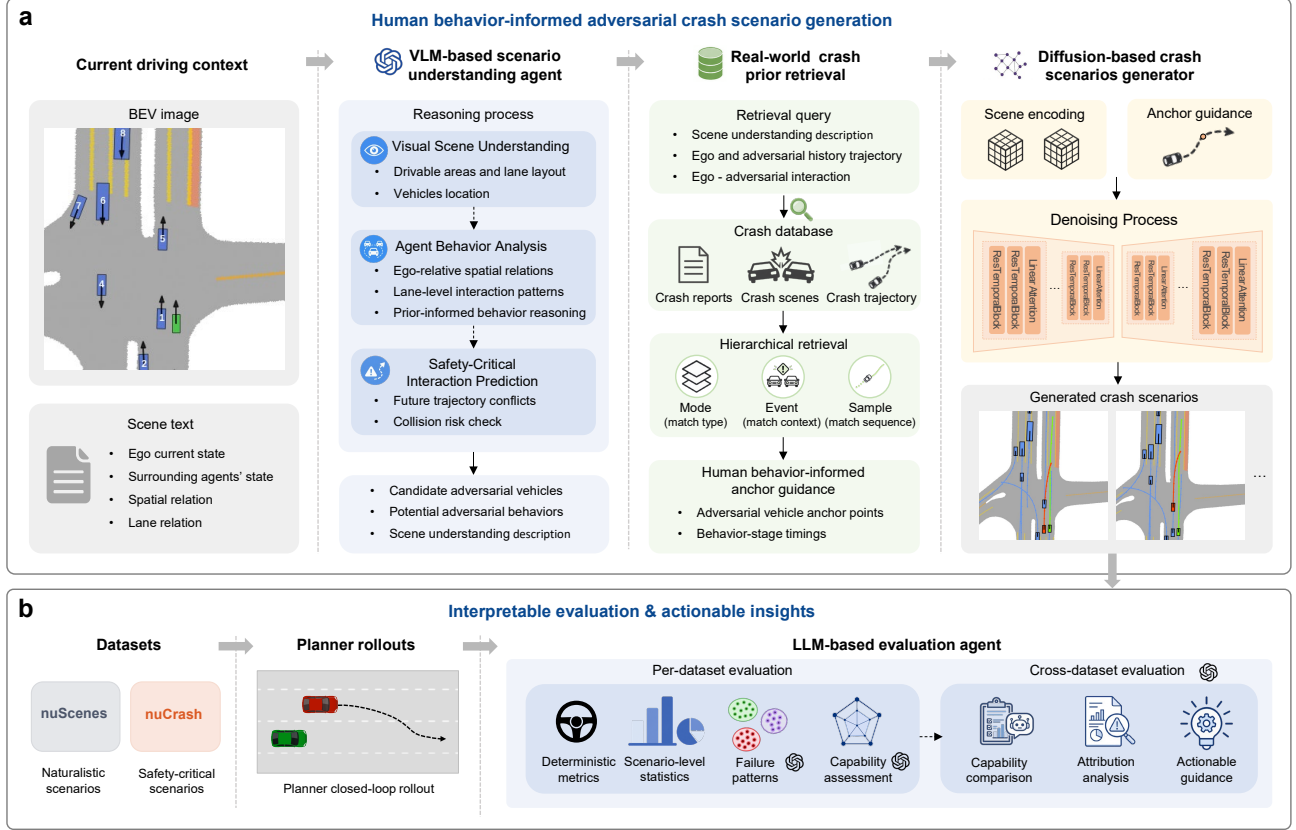


Figure 8. Human behavior-informed crash generation and interpretable safety evaluation with CrashSim. **a** Human behavior-informed adversarial crash scenario generation. CrashSim first interprets the current driving context through VLM-based scene understanding and retrieves relevant real-world crash priors through hierarchical matching. The retrieved crash prior provides human behavior-informed anchor guidance for a diffusion-based generator to produce safety-critical crash scenarios. **b** Interpretable evaluation and actionable insights. Generated crash scenarios form the long-tail nuCrash dataset which complements the naturalistic nuScenes dataset for closed-loop planner evaluation. An LLM-based evaluation agent performs both per-dataset and cross-dataset evaluation to diagnose failure mechanisms, assess planner capabilities, and provide actionable guidance for planner improvement.

crash priors, and generative multi-agent simulation to introduce realistic crash processes into naturalistic driving contexts. As illustrated in Fig. 8a, CrashSim takes a naturalistic driving scene as input, represented by a bird’s-eye-view (BEV) image and a structured textual description of the current driving context. A vision language model (VLM)-based scenario understanding agent interprets the scene and identifies candidate vehicles and potential behaviors that may develop into safety-critical interactions. Based on the scene understanding context, the historical trajectories of the ego vehicle and the VLM-identified candidate adversarial vehicle and potential adversarial behavior, and their historical interactions, CrashSim retrieves relevant real-world crash priors through a hierarchical retrieval process. The retrieved prior is then translated into human behavior-informed anchor guidance for a diffusion-based multi-agent trajectory generator, enabling the generation of realistic crash scenarios. This retrieval-augmented generation strategy grounds scenario generation in behavior and motion patterns observed in real-world crashes.

CrashSim further supports interpretable safety evaluation of autonomous driving planners, as shown in Fig. 8b. Planners are evaluated in closed loop on both naturalistic nuScenes scenarios and safety-critical nuCrash scenarios. An LLM-based evaluation agent analyzes deterministic metrics and scenario-level evidence across the two datasets to assess planner capabilities, diagnose recurrent failure mechanisms, and provide actionable guidance for planner improvement.

Problem formulation

We consider safety-critical multi-agent scenario generation for autonomous driving planner evaluation. Each scenario contains N vehicles, comprising an ego vehicle controlled by a planner π and $N - 1$ surrounding vehicles whose future motions are

generated by CrashSim. Among the surrounding vehicles, one is identified as the candidate adversarial vehicle that may develop a safety-critical interaction with the ego vehicle, whereas the others evolve as non-adversarial traffic participants.

A driving scenario is represented by a road map $\mathbf{m}$ and the states of all vehicles. At timestep t , the vehicle states are denoted as $\mathbf{s}_t = [s_t^0, s_t^1, \dots, s_t^{N-1}]$, where $s_t^i = (x_t^i, y_t^i, \theta_t^i, v_t^i)$ describes the two-dimensional position, heading and speed of vehicle i , and $i = 0$ denotes the ego vehicle. The historical trajectories over the previous T_{hist} timesteps are represented as $\mathbf{x} = \{\mathbf{s}_{t-T_{\text{hist}}}, \dots, \mathbf{s}_t\}$. The planner π produces the future trajectory of the ego vehicle over a horizon T , denoted as $\tau_{t:t+T}^0 = \pi(\mathbf{m}, \mathbf{x})$.

CrashSim further represents the current driving context using a BEV image $\mathbf{v}$ and a structured textual scene description $\mathbf{n}$. A VLM-based scenario understanding agent processes these inputs as $\mathbf{o} = \phi(\mathbf{v}, \mathbf{n})$, where $\mathbf{o} = (i^*, b, u)$ specifies the candidate adversarial vehicle i^* , its potential adversarial behavior b , and a structured textual scene understanding description u . Based on this output, CrashSim constructs a structured retrieval query $\mathbf{q} = (u, b, \mathbf{r}^{0,i^*}, \mathbf{x}^0, \mathbf{x}^{i^*})$, which integrates the scene understanding, predicted interaction behavior, pairwise interaction information, and historical trajectories of the ego and candidate adversarial vehicles. The retrieval process ρ searches the real-world crash knowledge base $\mathbf{d}_{\text{crash}}$ and returns a matched crash prior, $\mathbf{k} = \rho(\mathbf{q}, \mathbf{d}_{\text{crash}})$, which is subsequently transformed into human behavior-informed anchor guidance $\mathbf{a}$ for scenario generation.

Conditioned on the road map $\mathbf{m}$, historical trajectories $\mathbf{x}$ and anchor guidance $\mathbf{a}$, a latent diffusion model g_θ generates the future trajectories of the surrounding vehicles, $\tau_{t+T} = g_\theta(\mathbf{m}, \mathbf{x}, \mathbf{a})$. The model comprises an encoder E that maps the scene context into a latent representation and a decoder D that reconstructs the generated trajectories from the latent space. For constructing the nuCrash dataset used in this study, the ego vehicle follows its recorded future trajectory in the corresponding nuScenes scene. The resulting scenarios are retained as a fixed evaluation set shared across all planners. During subsequent closed-loop planner evaluation, the recorded ego trajectory is replaced by the trajectory produced by the evaluated planner π , which is replanned every 0.5 s.

VLM-based scenario understanding agent

The VLM-based scenario understanding agent interprets the current driving context and identifies surrounding vehicles that may plausibly develop safety-critical interactions with the ego vehicle. It jointly processes the BEV image, which captures drivable areas, lane geometry, and vehicle locations and headings, and structured scene text describing vehicle states and ego-relative spatial and lane relations. The VLM-based scenario understanding agent is implemented using GPT-4o⁴⁷.

The reasoning process comprises three stages. First, *visual scene understanding* identifies drivable regions, lane layout and connectivity, and the lane-level locations of surrounding vehicles, thereby establishing the geometric constraints on feasible future motion. Second, *agent behavior analysis* reasons about ego-relative interactions and plausible surrounding vehicle behaviors. Real-world crash-pattern statistics are incorporated as prior knowledge to guide behavior inference and candidate selection, while scene geometry and observed motion remain the primary evidence. Third, *safety-critical interaction prediction* assesses whether the inferred future motions may create trajectory conflicts or insufficient safety margins with the ego vehicle. A feasibility check removes candidates that are unlikely to interact with the ego vehicle within the prediction horizon. The resulting candidate adversarial vehicles, potential adversarial behaviors, and structured scene-understanding description are used to construct the query for real-world crash-prior retrieval. Detailed VLM prompts, together with an example BEV image and its corresponding structured scene text, are provided in Supplementary Sections.

Real-world crash-prior retrieval and anchor construction

Database construction. We constructed a structured real-world crash database from bird's-eye-view trajectory reconstructions of naturalistic crashes and near-crashes in the SHRP2 NDS²⁹. Each retained safety-critical interaction was assigned to one of 16 interaction modes derived from the NHTSA pre-crash scenario typology²⁷. The trajectories were transformed into an ego-centred coordinate frame and resampled at 0.5 s intervals. The processed interactions were organized into a three-layer database covering interaction semantics, individual crash events and temporal motion patterns. The mode layer represents each interaction class through its semantic description, vehicle manoeuvres and contextual factors. The event layer represents individual safety-critical events using event descriptions and event-level motion characteristics. The sample layer decomposes each event into 8 s windows. The first 2 s describe the observed motion, while the subsequent 6 s target-vehicle trajectory provides the motion anchor for downstream generation. Moreover, natural-language descriptors at all three layers were encoded using the all-MiniLM-L6-v2 Sentence-Transformer model^{48,49}, providing a consistent semantic representation of the textual information associated with the real-world crash records.

The paired ego-target trajectories within the final 6 s of each 8 s window were automatically annotated with behavior-stage timings using a rule-based strategy. We first define a kinematic conflict condition as

$$R_j(\tau, \gamma, \alpha) = (\text{TTC}_j \leq \tau) \vee (c_j \geq \gamma) \vee (d_j \leq \alpha d_0), \quad (1)$$

where d_j , c_j and TTC_j denote relative distance, closing speed and time to collision (TTC) at time step j , respectively. The onset indices of the risk-awareness, emergency-response and collision stages, denoted by t_{risk} , t_{emg} and t_{col} , respectively, are

then defined as

$$\begin{aligned} t_{\text{risk}} &= \min \{ j < t_{\text{col}} : R_j(\tau_r, \gamma_r, \alpha_r) \}, \\ t_{\text{emg}} &= \min \{ j < t_{\text{col}} : [R_j(\tau_e, \gamma_e, \alpha_e) \wedge a_j \leq -\beta_e] \vee a_j \leq -\beta_s \}, \\ t_{\text{col}} &= \min \{ j : g_j \leq \varepsilon \}. \end{aligned} \quad (2)$$

Here, a_j is the target vehicle's longitudinal acceleration, g_j is the minimum separation between the oriented bounding boxes, and d_0 is the initial relative distance. The positive parameters β_e and β_s denote the magnitudes of the emergency and strong-deceleration thresholds, respectively, with $\beta_s > \beta_e > 0$. The parameters τ , γ and α specify the TTC, closing-speed and relative-distance thresholds. If no qualifying pre-collision time step existed, the corresponding stage onset was recorded as undefined. Details of the database construction are provided in Supplementary Section.

Prior retrieval and anchor construction. CrashSim implements the retrieval operator ρ as a three-stage cascade over the crash database $\mathbf{d}_{\text{crash}}$. Mode retrieval first identifies interaction classes that are semantically compatible with the query $\mathbf{q}$. Event retrieval then ranks individual events according to their interaction characteristics and motion consistency. Sample retrieval finally identifies the temporal motion pattern that best matches the observed motion.

Text embeddings are compared using cosine similarity, consistent with the Sentence-Transformer framework⁴⁸. Cosine similarity is also applied to the signed event-feature and differential-motion vectors, with the resulting scores linearly rescaled to $[0, 1]$ before fusion. Categorical, state and spatial similarities were normalized separately and combined using layer-specific weights.

Among the sample candidates retained after mode and event retrieval, CrashSim selects the candidate with the highest composite sample-level similarity score S_{sample} . The matched crash prior $\mathbf{k} = \rho(\mathbf{q}, \mathbf{d}_{\text{crash}})$ is retained only when this score exceeds a predefined threshold δ ; otherwise, no crash-derived anchor is passed to the downstream trajectory generator.

The accepted crash prior $\mathbf{k}$ contains a 6 s differential-motion sequence,

$$\mathbf{k}_{\text{motion}} = \{(\Delta \ell_j, \Delta \theta_j)\}_{j=1}^F, \quad F = 12, \quad (3)$$

where $\Delta \ell_j$ and $\Delta \theta_j$ denote the distance increment and heading change over the j -th 0.5-s interval, respectively. Rather than directly transferring the absolute trajectory of the source crash, CrashSim preserves its relative-motion pattern and reconstructs the anchor trajectory from the current state s_t^* using midpoint integration:

$$\tilde{\theta}_{j-\frac{1}{2}} = \tilde{\theta}_{j-1} + \frac{1}{2} \Delta \theta_j, \quad \tilde{p}_j = \tilde{p}_{j-1} + \Delta \ell_j \begin{bmatrix} \cos \tilde{\theta}_{j-\frac{1}{2}} \\ \sin \tilde{\theta}_{j-\frac{1}{2}} \end{bmatrix}, \quad \tilde{\theta}_j = \text{wrap}(\tilde{\theta}_{j-1} + \Delta \theta_j), \quad (4)$$

where $\tilde{p}_0 = (x_t^*, y_t^*)^\top$ and $\tilde{\theta}_0 = \theta_t^*$. The reconstructed positions form the crash-informed anchor trajectory $\mathbf{P}_{\text{anc}} = (\tilde{p}_1, \dots, \tilde{p}_F)$.

The behavior-stage timings associated with the matched crash prior $\mathbf{k}$ are retrieved together with its motion pattern and collected as $\mathbf{t}_{\text{stage}} = (t_{\text{risk}}, t_{\text{emg}}, t_{\text{col}})$. The resulting anchor guidance is represented as

$$\mathbf{a} = (i^*, \mathbf{P}_{\text{anc}}, \mathbf{t}_{\text{stage}}). \quad (5)$$

This construction preserves the temporal motion pattern and behavior-stage progression of the matched real-world crash while spatially aligning the reconstructed anchor trajectory with the current state of the candidate adversarial vehicle. Further implementation details are provided in Supplementary Section.

Diffusion-based crash scenario generation

CrashSim uses a scene-conditioned latent diffusion model to generate the future trajectories $\tau_{t:t+T}^{1:N-1}$ of the surrounding vehicles. Given the road map and observed multi-agent motion history $(\mathbf{m}, \mathbf{x})$, the encoder E produces a scene-conditioning latent $\mathbf{c} = E(\mathbf{m}, \mathbf{x})$. Both the encoder and decoder adopt interaction-aware graph architectures to model dependencies among traffic participants^{15, 19}. The denoising network adopts a temporal U-Net composed of residual temporal blocks and linear-attention layers⁵⁰. The diffusion model is trained on the nuScenes training split using the standard noise-prediction objective⁵¹, with the denoising network conditioned on the scene context $\mathbf{c}$. Model architecture and training details are provided in Supplementary Section.

At inference, sampling is initialized from Gaussian noise, $\mathbf{z}^K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and progressively denoises the latent through the DDIM reverse process⁵²:

$$\mathbf{z}^{k-1} = \sqrt{\bar{\alpha}_{k-1}} \left(\frac{\mathbf{z}^k - \sqrt{1 - \bar{\alpha}_k} \varepsilon_\theta(\mathbf{z}^k, k, \mathbf{c})}{\sqrt{\bar{\alpha}_k}} \right) + \sqrt{1 - \bar{\alpha}_{k-1}} \varepsilon_\theta(\mathbf{z}^k, k, \mathbf{c}), \quad (6)$$

where $\bar{\alpha}_k$ denotes the cumulative noise schedule and ε_θ denotes the scene-conditioned denoising network that predicts the noise component at diffusion step k . The reverse process iteratively removes noise from $\mathbf{z}^K$ to obtain the clean latent $\mathbf{z}^0$. Finally, the decoder D maps $\mathbf{z}^0$, together with the scene context, to the joint future trajectories $\tau_{t:t+T}^{1:N-1}$ of the surrounding vehicles. The decoder further incorporates autoregressive motion prediction and kinematic-bicycle-model propagation to produce physically continuous and dynamically feasible multi-agent trajectories¹⁵.

A key component of CrashSim is *human behavior-informed anchor guidance*, which integrates the temporal interaction structure of real-world crashes into the generative process to facilitate the generation of realistic crash trajectories. The anchor $\mathbf{a}$ comprises spatial waypoints and behavior-stage annotations, which jointly guide the diffusion process. This representation preserves characteristic human responses preceding and during a crash while allowing the generated trajectory to adapt to the geometry and traffic configuration of the current scene.

During the final few reverse diffusion steps, we apply guidance based on the clean-guidance strategy⁵³. At each guided step, the predicted clean latent $\hat{\mathbf{z}}^0$ is decoded into trajectory space, where a differentiable guidance loss J is computed based on the decoded trajectories and the anchor guidance $\mathbf{a}$. Following gradient-based diffusion guidance⁵⁴, the noise prediction is modified as

$$\tilde{\varepsilon}_\theta(\mathbf{z}^k, k, \mathbf{c}) = \varepsilon_\theta(\mathbf{z}^k, k, \mathbf{c}) - s \nabla_{\mathbf{z}^k} J[D_\theta(\hat{\mathbf{z}}^0, \mathbf{x}, \mathbf{m}); \mathbf{a}], \quad (7)$$

where s denotes the guidance scale. During the guided steps, $\tilde{\varepsilon}_\theta$ replaces ε_θ in the DDIM update in Eq. (6) to obtain $\mathbf{z}^{k-1}$.

The guidance loss combines behavioral anchoring, stage consistency, safety-critical interaction, trajectory realism and route consistency:

$$J = w_{\text{anchor}} J_{\text{anchor}} + w_{\text{stage}} J_{\text{stage}} + w_{\text{coll}} J_{\text{coll}} + w_{\text{real}} J_{\text{real}} + w_{\text{route}} J_{\text{route}}, \quad (8)$$

where the weights control the relative contributions of the five guidance terms.

The anchor guidance term transfers the spatial progression and response dynamics of the retrieved crash anchor to the candidate adversarial vehicle i^* :

$$J_{\text{anchor}} = \left\langle w_p \rho(\|p_t - \bar{p}_t\|_2) + w_a (a_t - \bar{a}_t)^2 + w_\psi \text{wrap}(\Delta\psi_t - \Delta\bar{\psi}_t)^2 \right\rangle_{t \in \mathcal{V}}, \quad (9)$$

where p_t , a_t and $\Delta\psi_t$ denote position, longitudinal acceleration and heading change, respectively. Barred quantities are computed from the anchor, and ρ denotes the smooth- L_1 penalty. The set $\mathcal{V}$ contains pre-impact anchor waypoints and excludes off-road segments occurring after the anchor has entered the drivable area. This treatment preserves supervision at initial spawn locations while preventing incompatible post-entry segments from being reproduced.

The stage guidance term aligns the generated interaction with the risk-awareness, emergency-response and collision stages encoded by the retrieved anchor. Using the stage timings defined above, let t_{risk} , t_{emg} and t_{col} denote the corresponding timings inferred from the generated interaction, while t_{risk}^* , t_{emg}^* and t_{col}^* denote those of the anchor. The stage objective is

$$J_{\text{stage}} = w_r (t_{\text{risk}} - t_{\text{risk}}^*)^2 + w_m (t_{\text{emg}} - t_{\text{emg}}^*)^2 + w_c (t_{\text{col}} - t_{\text{col}}^*)^2 + w_d E_{\text{dur}} + w_o E_{\text{ord}}, \quad (10)$$

The first three terms align the onset of risk awareness, emergency response and collision with the retrieved crash anchor, while E_{dur} discourages shortening of the emergency-response-to-collision interval and E_{ord} penalizes violations of the expected risk-awareness-emergency-response-collision ordering. In implementation, the timing and duration errors are normalized by the prediction horizon to ensure a consistent scale across trajectories, and only terms with valid stage annotations are included. The stage timings are obtained using differentiable soft versions of the behavior-stage criteria defined above, with the detailed normalization and weighting coefficients provided in the Supplementary Section.

Together, J_{anchor} and J_{stage} transfer human response patterns from real-world crash processes into the generated interaction.

The collision term J_{coll} promotes a safety-critical interaction between adversarial vehicle i^* and the ego vehicle, while the route term J_{route} encourages i^* to follow a plausible route consistent with the reference motion pattern. The realism term J_{real} penalizes off-road motion and unintended collisions. Following LD-scene¹⁹, role-specific gradient masks apply all five guidance terms to the adversarial vehicle i^* , while only the realism term J_{real} is applied to the remaining surrounding vehicles; no gradient is applied to the ego trajectory. Overall, this guidance formulation preserves variation in interaction timing, conflict geometry and vehicle responses while retaining the behavioral characteristics of real-world crash processes.

For each scene-anchor pair, CrashSim samples multiple latent realizations and retains the candidate with the lowest guidance score after penalizing violations of physical feasibility. Details of guidance terms are provided in Supplementary Section.

LLM-assisted planner evaluation

We develop an LLM-assisted evaluation framework to characterize how planner capabilities change from naturalistic driving in nuScenes to safety-critical interactions in nuCrash (Fig. 8b). nuCrash is generated by CrashSim to introduce human behavior-informed safety-critical interactions while preserving the scene and trajectory representation used for nuScenes. Each planner is evaluated on both datasets under the same closed-loop rollout protocol, with its behavior quantified using deterministic metrics and scenario-level statistics covering safety, response timing, longitudinal and lateral control, comfort and stability, and mobility efficiency.

While these measurements quantify individual aspects of planner performance, they remain fragmented across metrics, scenarios and failure patterns and do not directly reveal the underlying capability limitations. We therefore introduce an LLM-assisted evaluation agent, implemented using GPT-4o⁴⁷, to analyze this heterogeneous evidence through evidence analysis and structured reasoning^{33,34}. For each dataset, the agent jointly interprets metric outcomes, failure patterns and their scenario contexts to derive a capability profile along six dimensions: risk anticipation, interaction reasoning, longitudinal response, lateral response, comfort and stability, and mobility efficiency. This analysis identifies consistent behavioral limitations and the interaction conditions under which they emerge, rather than relying on any individual metric as a direct indicator of planner capability.

The evaluation agent then performs cross-dataset analysis by contrasting the capability profiles obtained from nuScenes and nuCrash. This comparison distinguishes limitations already present in naturalistic driving from those amplified or newly exposed under safety-critical interactions. By relating capability changes to the corresponding metric shifts, failure patterns and scenario contexts, the agent identifies the associated failure mechanisms, attributes them to underlying behavioral limitations, and translates the resulting diagnosis into targeted directions for planner improvement and subsequent validation. Detailed metric definitions and prompts are provided in Supplementary Section.

References

1. Tafidis, P., Farah, H., Brijs, T. & Pirdavani, A. Safety implications of higher levels of automated vehicles: a scoping review. *Transp. reviews* **42**, 245–267 (2022).
2. Sohrabi, S., Khodadadi, A., Mousavi, S. M., Dadashova, B. & Lord, D. Quantifying the automated vehicle safety performance: A scoping review of the literature, evaluation of methods, and directions for future research. *Accid. Analysis & Prev.* **152**, 106003 (2021).
3. Nunes, A., Reimer, B. & Coughlin, J. F. People must retain control of autonomous vehicles. *Nature* **556**, 169–171 (2018).
4. Liu, H. X. & Feng, S. Curse of rarity for autonomous vehicles. *Nat. Commun.* **15**, 4808 (2024).
5. Njord, J. & Steudle, K. Big data hit the road: The first year of use of the SHRP 2 safety databases. *TR News* 3–8 (2015).
6. Kalra, N. & Paddock, S. M. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transp. Res. Part A: Policy Pract.* **94**, 182–193 (2016).
7. Ding, W. *et al.* A survey on safety-critical driving scenario generation—a methodological perspective. *IEEE Transactions on Intell. Transp. Syst.* **24**, 6971–6988 (2023).
8. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A. & Koltun, V. CARLA: An open urban driving simulator. In *the Proceedings of Conference on Robot Learning (CoRL)*, 1–16 (2017).
9. Lopez, P. A. *et al.* Microscopic traffic simulation using sumo. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2575–2582 (2018).
10. Feng, S., Yan, X., Sun, H., Feng, Y. & Liu, H. X. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nat. Commun.* **12**, 748 (2021).
11. Feng, S. *et al.* Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* **615**, 620–627 (2023).
12. Qian, C., Xu, J., Xing, X. & Guo, F. Test case sampling optimization for safety validation of automated driving systems. *Nat. Commun.* **17**, 3114 (2026).
13. Yan, X. *et al.* Learning naturalistic driving environment with statistical realism. *Nat. Commun.* **14**, 2037 (2023).
14. Wang, J. *et al.* AdvSim: Generating safety-critical scenarios for self-driving vehicles. In *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9909–9918 (IEEE, 2021).
15. Rempe, D., Philion, J., Guibas, L. J., Fidler, S. & Litany, O. Generating useful accident-prone driving scenarios via a learned traffic prior. In *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17284–17294 (IEEE, 2022).
16. Hanselmann, N., Renz, K., Chitta, K., Bhattacharyya, A. & Geiger, A. KING: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In *the Proceedings of European Conference on Computer Vision (ECCV)*, 335–352 (Springer, 2022).
17. Chang, W.-J., Pittaluga, F., Tomizuka, M., Zhan, W. & Chandraker, M. SAFE-SIM: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries. In *the Proceedings of European Conference on Computer Vision (ECCV)*, 242–258 (Springer, 2024).
18. Peng, M. *et al.* Safety-critical traffic simulation with guided latent diffusion model. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 987–993 (IEEE, 2025).
19. Peng, M. *et al.* LD-scene: LLM-guided diffusion for controllable generation of adversarial safety-critical driving scenarios. *Transp. Res. Part C: Emerg. Technol.* **188**, 105694 (2026).
20. Wu, X., Chen, J., Zhong, N. & Shen, Y. VLM as strategist: Adaptive generation of safety-critical testing scenarios via guided diffusion. *Accid. Analysis & Prev.* **236**, 108680 (2026).
21. Mei, Y., Nie, T., Sun, J. & Tian, Y. LLM-Attacker: Enhancing closed-loop adversarial scenario generation for autonomous driving with large language models. *IEEE Transactions on Intell. Transp. Syst.* **26**, 15068–15076 (2025).
22. Jiang, Z., Li, Z., Wang, C., Wang, Z. & Xiong, C. AnchorDrive: Llm scenario rollout with anchor-guided diffusion regeneration for safety-critical scenario generation. *arXiv preprint arXiv:2603.02542* <https://arxiv.org/abs/2603.02542> (2026).
23. Chang, C., Wang, S., Zhang, J., Ge, J. & Li, L. LLMscenario: Large language model driven scenario generation. *IEEE Transactions on Syst. Man, Cybern. Syst.* **54**, 6581–6594 (2024).

24. Svärd, M., Markkula, G., Bärghman, J. & Victor, T. Computational modeling of driver pre-crash brake response, with and without off-road glances: Parameterization using real-world crashes and near-crashes. *Accid. Analysis & Prev.* **163**, 106433 (2021).
25. Guo, H., Xie, K. & Keyvan-Ekbatani, M. Modeling driver's evasive behavior during safety-critical lane changes: Two-dimensional time-to-collision and deep reinforcement learning. *Accid. Analysis & Prev.* **186**, 107063 (2023).
26. Schumann, J. F. *et al.* Active inference as a model of collision avoidance behavior in human drivers. *Nat. Commun.* **17**, 5009 (2026).
27. Najm, W. G., Smith, J. D. & Yanagisawa, M. Pre-crash scenario typology for crash avoidance research. Tech. Rep. DOT HS 810 767, National Highway Traffic Safety Administration (2007).
28. Hankey, J. M., Perez, M. A. & McClafferty, J. A. Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets. Tech. Rep. SHRP2-S2-S06-RR-1, Virginia Tech Transportation Institute (2016).
29. Jiao, Y., Calvert, S. C., van Cranenburgh, S. & van Lint, H. Learning collision risk proactively from naturalistic driving data at scale. *Nat. Mach. Intell.* **8**, 337–350 (2026).
30. Fang, J., Yan, D., Qiao, J., Xue, J. & Yu, H. DADA: Driver attention prediction in driving accident scenarios. *IEEE Transactions on Intell. Transp. Syst.* **23**, 4959–4971 (2021).
31. Yao, Y. *et al.* DoTA: Unsupervised detection of traffic anomaly in driving videos. *IEEE Transactions on Pattern Analysis Mach. Intell.* **45**, 444–459 (2022).
32. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. In *the Proceedings of Advances in Neural Information Processing Systems*, vol. 33, 9459–9474 (2020).
33. Shi, X., Zhao, W., Chen, T., Yang, C. & Du, J. Evidence triangulator: using large language models to extract and synthesize causal evidence across study designs. *Nat. Commun.* **16**, 7355 (2025).
34. Qiu, P. *et al.* Quantifying the reasoning abilities of llms on clinical cases. *Nat. Commun.* **16**, 9799 (2025).
35. Caesar, H. *et al.* NuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810* <https://arxiv.org/abs/2106.11810> (2021).
36. Peng, M., Yao, R., Guo, X. & Ma, J. nuPlan-R: A closed-loop planning benchmark for autonomous driving via reactive multi-agent simulation. In *IEEE Intelligent Vehicles Symposium*, 702–709 (IEEE, 2026).
37. Caesar, H. *et al.* nuScenes: A multimodal dataset for autonomous driving. In *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11621–11631 (IEEE, 2020).
38. Antin, J. F. *et al.* Second strategic highway research program naturalistic driving study methods. *Saf. Sci.* **119**, 2–10 (2019).
39. Li, T. *et al.* World Engine: Towards the era of post-training for autonomous driving. *arXiv preprint arXiv:2606.19836* <https://arxiv.org/abs/2606.19836> (2026).
40. Gibbs, A. L. & Su, F. E. On choosing and bounding probability metrics. *Int. Stat. Rev.* **70**, 419–435 (2002).
41. Lin, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Inf. Theory* **37**, 145–151 (1991).
42. Montemerlo, M. *et al.* Junior: The stanford entry in the urban challenge. *J. Field Robotics* **25**, 569–597 (2008).
43. Dauner, D., Hallgarten, M., Geiger, A. & Chitta, K. Parting with misconceptions about learning-based vehicle motion planning. In *the Proceedings of Conference on Robot Learning (CoRL)*, 1268–1281 (2023).
44. Bansal, M., Krizhevsky, A. & Ogale, A. ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079* <https://arxiv.org/abs/1812.03079> (2018).
45. Suo, S., Regalado, S., Casas, S. & Urtasun, R. TrafficSim: Learning to simulate realistic multi-agent behaviors. In *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10400–10409 (IEEE, 2021).
46. Zheng, Y. *et al.* Diffusion-based planning for autonomous driving with flexible guidance. In *the Proceedings of International Conference on Learning Representations (ICLR)*, vol. 2025, 37207–37227 (2025).
47. Hurst, A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* <https://arxiv.org/abs/2410.21276> (2024).
48. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992 (2019).

49. Wang, W. *et al.* MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Proc. Adv. Neural Inf. Process. Syst.* **33**, 5776–5788 (2020).
50. Janner, M., Du, Y., Tenenbaum, J. B. & Levine, S. Planning with diffusion for flexible behavior synthesis. In *the Proceedings of International Conference on Machine Learning (ICML)*, 9902–9915 (2022).
51. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Proc. Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
52. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* <https://arxiv.org/abs/2010.02502> (2020).
53. Remppe, D. *et al.* Trace and Pace: Controllable pedestrian animation via guided trajectory diffusion. In *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13756–13766 (IEEE, 2023).
54. Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis. *Proc. Adv. Neural Inf. Process. Syst.* **34**, 8780–8794 (2021).