Beyond Binary Metrics: Unveiling the Safety Illusion in Autonomous Driving Simulation

Zexin Feng Linyu Xiao Xintao Yan*
Department of Civil Engineering, The University of Hong Kong
{zexinfeng, lingyu.xiao}@connect.hku.hk, xintaoy@hku.hk

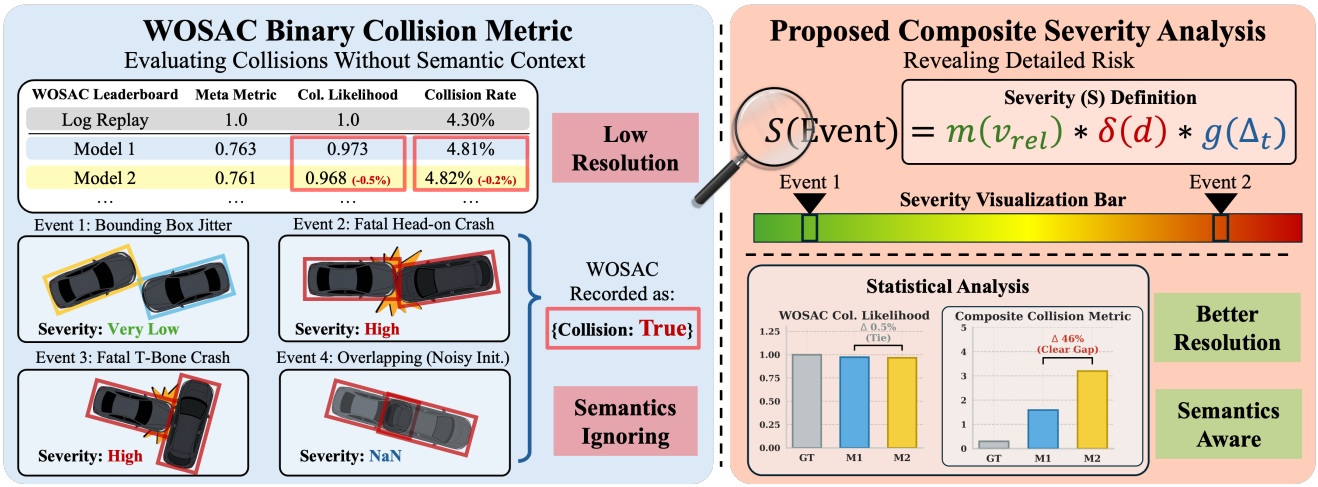

Figure 1. **Breaking the Safety Illusion.** Widely used traffic simulation benchmarks (e.g., the Waymo Open Sim Agent Challenge, (WOSAC)) treat all collisions equally (left), collapsing minor contacts and catastrophic crashes into the same binary event. Our **Composite Collision Metric (CCM)** (right) introduces severity-aware analysis that distinguishes fatal crashes from minor jitters, enabling more accurate and reliable evaluation of simulation fidelity for training and benchmarking traffic simulators.

Abstract

Long-horizon closed-loop simulation is increasingly used to evaluate and improve autonomous driving systems, yet its safety realism is still largely judged by benchmark-facing binary collision metrics. We argue that this creates a safety illusion*: simulators can achieve near-saturated benchmark scores while still exhibiting substantially different physical risk. We trace this limitation to a fundamental* Resolution Gap *in current evaluation, where collision outcomes are collapsed into rollout-level binary indicators that capture event frequency but not failure severity. To address this problem, we propose* Composite Severity Analysis*, a physically grounded framework that decomposes pairwise collisions into relative impact velocity, penetration depth, and contact duration, and aggregates them into a continuous severity representation with taxonomy-aware noise filtering. Building on this formulation, we introduce the* Composite Collision Metric (CCM)*, which jointly captures collision occurrence and consequence severity. Experiments on Waymo Open Sim Agent Challenge (WOSAC) show that simulators with nearly identical benchmark-facing collision performance can exhibit sharply different tail-risk profiles, while safety-oriented improvements that are nearly invisible to standard metrics become clearly measurable under CCM. These results suggest that binary collision likelihood is insufficient for simulator safety evaluation and that severity-aware analysis is necessary for trustworthy safety alignment in autonomous driving simulation.*

1. Introduction

To bridge the gap between offline training and real-world deployment, long-horizon and high-fidelity closed-loop simulation has become an indispensable proxy for evalu-

*Corresponding author.

ating policy robustness and enabling iterative safety alignment [20, 29, 41, 47, 48]. The autonomous driving community increasingly relies on these simulation platforms for training and testing autonomous agents in a realistic virtual environment prior to deployment[16, 17, 30, 47, 51]. This demand has spurred the development of advanced multi-agent simulation systems [44, 50, 53], which have achieved impressive performance on standardized benchmarks such as the Waymo Open Sim Agent Challenge (WOSAC) [31]. However, while these simulators are expected to guarantee safe autonomous behaviors, a paradoxical challenge arises: the intrinsic safety realism of the simulation is not guaranteed.

As generative sim-agents progressively saturate benchmark-facing safety metrics, a critical evaluation bottleneck emerges. We identify a fundamental Resolution Gap rooted in the mathematical formulation of current collision metrics. In standards such as WOSAC, collision is primarily modeled as a binary rollout-level indicator aligned with the Ground Truth (GT) distributions of the Waymo Open Motion Dataset (WOMD) [15, 31]. More broadly, existing safety evaluation in traffic simulation and motion forecasting often relies on coarse event-level surrogates, such as collision frequency, binary collision labels, or time-to-collision-style indicators, which provide limited resolution on the physical consequence of failure [10, 18]. By rewarding sim-agents to match 0/1 event likelihoods observed in the dataset, such scoring protocols primarily constrain marginal event frequency rather than the conditional severity structure of collisions. Recent literature and community discussions have similarly highlighted this limitation, noting that benchmark-oriented distribution matching can reward reproducing sensor noise, annotation artifacts, or anomalous behaviors rather than genuinely safe or human-like dynamics [6, 11].

As shown in Fig. 1, this coarse representation leads to a critical loss of physical nuance, completely masking the realism of long-tail dangerous events. Under a binary lens, a 0.1-second minor grazing contact, which is often a byproduct of bounding-box jitter or sensor noise, is encoded as identical to a fatal, high-speed T-bone crash. Both behaviors collapse into the same "collision=1" state, rendering the alignment process becomes indifferent to underlying safety semantics. It provides no mathematical incentive for the model to distinguish between nominal data artifacts and catastrophic failures, fundamentally ignoring the physical realism of these rare but critical tail events as long as the aggregate likelihood matches the metadata.

This indifference gives rise to a "safety illusion": a phenomenon in which simulation systems achieve near-perfect scores under benchmark evaluation, yet the simulation itself lacks any guarantee of intrinsic safety realism. Conceptually, this limitation is consistent with a broader literature showing that proxy-aligned objectives can fail to preserve the semantic property that practitioners care about[28, 32], especially in the presence of rare but consequential failures. For a world model to reliably serve as a trusted environment for both training (e.g., policy refinement via Reinforcement Learning) and testing autonomous driving systems[3, 37], its evaluation signal must provide monotonic, physically grounded feedback. A binary metric, bounded by step-function saturation and devoid of severity awareness, fails to provide the informative gradients necessary to penalize severe anomalies, allowing compounding errors to cascade into systemic failure during long-horizon closed-loop simulations[7, 36].

In this paper, we propose Composite Severity Analysis, a physically grounded evaluation framework that recaptures the lost safety semantics. Our core contributions are:

- **Analytical Framework of the Safety Illusion**: We show that binary collision likelihood creates a safety illusion in the high-score regime of modern traffic simulators. By formalizing the non-identifiability inherent in binary collision metrics, we demonstrate that matching aggregate collision likelihoods fundamentally masks fatal long-tail kinematic risks.
- **Empirical Revelation of Hidden Tail Risks**: Using state-of-the-art simulators, we reveal that models with nearly identical WOSAC collision scores exhibit radically divergent long-tail risks (e.g., Conditional Value-at-Risk CVaR_α), proving that binary likelihood is an insufficient target for safety alignment.
- **The Composite Collision Metric (CCM)**: We propose a continuous, physically grounded evaluation framework that provides the high-resolution signal necessary for optimization. We empirically demonstrate that CCM effectively isolates structural damage from simulation noise, offering a highly discriminative landscape for safety-critical policy iteration.

2. Related Works

2.1. From Planning Evaluation to the Need for High-Fidelity Simulation

The evaluation metric largely depends on benchmarks. Early works on end-to-end planning, such as UniAD and TransFuser [8, 23], have a preliminary focus on open-loop metrics like MSE or MAE. To address the limitations of open-loop evaluation, other works, including VAD, ROACH, PlanT, and DriveAdaptor [24, 26, 34, 49] that test on CARLA [13], or Bench2Drive [25] formulate composite driving scores that heavily penalize safety infractions. Recent works on NAVSIM [12] utilized a modified PDM score. Moreover, there are also works [46] implemented on nuPlan [4] or Waymax [19] using metrics like drivable area compliance, kinematic feasibility, or arrival rate under

different driving scenarios [45, 46]. This evolution underscores a community consensus: rigorous closed-loop safety assessment is paramount for evaluating planning models. However, the reliability of these closed-loop metrics is not solely a property of the planner; it intrinsically depends on the physical fidelity of the underlying simulation environment.

2.2. The Rise of Generative Traffic Sim-Agents

Driven by the demand for high-fidelity closed-loop environments highlighted above, learning-based sim-agents have evolved rapidly. Tokenized autoregressive simulators such as Trajeglish, SMART, and BehaviorGPT [33, 44, 53] improve scalability and interaction modeling through map/trajectory tokenization and next-token prediction, enabling efficient closed-loop rollouts. CAT-K further boosts closed-loop performance via supervised fine-tuning to mitigate covariate shift. Recent advancements, such as SPACeR [6] and RLFTSim [1], have pushed benchmark limits even further through reinforcement learning (RL) fine-tuning. However, the success of these highly capable models inherently relies on the quality of the evaluation metrics used as their optimization targets, which exposes a critical vulnerability in the current simulation paradigm.

2.3. The Resolution Gap in Simulation Metrics

This reliance brings the evaluation metrics themselves into question. Evaluation in traffic simulation has evolved from open-loop displacement errors in motion forecasting (e.g., minADE for MTR/MTR++ [38, 39]) to early closed-loop heuristics like binary collision and off-road rates [13, 40]. To assess multi-agent interactive realism at scale, standardized protocols like WOSAC [31] introduced distributional matching, quickly becoming the primary evaluation target for the modern behavior models [1, 33, 44, 50]. However, despite this paradigm shift, their underlying safety components retain a fundamental flaw: they still treat infractions as binary or weakly aggregated events, offering zero resolution on physical consequence severity. As we demonstrate, in the high-score regime of modern generative simulators, this coarse granularity creates a safety illusion that severely masks catastrophic long-tail risks.

3. Preliminaries

To understand the necessity of a physically grounded severity metric, we must formalize how safety is currently evaluated in traffic simulation benchmarks[19] and why this formulation breaks down during policy iteration.

3.1. Formalizing the Metric Coarseness

Current protocols, such as WOSAC [31], operationalize collision likelihood through coarse code-level aggregations rather than physical semantics. Given that a single scenario can yield multiple rollout versions, the collision status is evaluated per agent within each rollout. For any agent i in a specific rollout, the simulator computes the minimum signed distance to *any* other object at each timestep t. A collision is flagged if this distance falls below a threshold, stripping away all contextual semantics at the moment of impact to register only the bare occurrence of the event. Critically, these step-wise flags are temporally compressed into a single binary indicator for that agent in that particular rollout:

$$C_i = 1[\exists t \in [0, T],\ \text{distance}_i(t) < \text{threshold}] \quad (1)$$

This aggregation fundamentally equates a minor, transient graze against a static obstacle with a severe, prolonged penetration involving multiple vehicles. By collapsing interaction frequency, contact duration, and structural depth into a simple "ever collided or not" boolean, the metric forces the evaluation to prioritize statistical binary matching over the mitigation of actual physical risk.

3.2. Mathematical Proof of the Safety Illusion

This binary reduction creates a fatally flawed optimization landscape. In the WOMD [15], nominal data artifacts and sensor noise yield a strictly non-zero ground-truth collision rate of roughly 4.2%:

$$P_{\text{data}}(C = 1) \approx 0.042 \quad (2)$$

Benchmark metrics evaluate realism by penalizing distributional divergence. Therefore, a simulator parameterized by θ maximizes its evaluation score $\mathcal{M}$ through strict statistical mimicry of this noise floor:

$$P_\theta(C = 1) \to P_{\text{data}}(C = 1) = 0.042 \quad (3)$$

Let S denote the true physical severity of an interaction. By the Law of Total Expectation, the expected physical risk of the deployed policy is:

$$\mathbb{E}_\theta[S] = P_\theta(C = 1) \cdot \mathbb{E}_\theta[S \mid C = 1] \quad (4)$$

Herein lies the mathematical iteration bottleneck. Because the benchmark score $\mathcal{M}$ depends exclusively on the marginal probability $P_\theta(C = 1)$, its gradient with respect to conditional physical severity is exactly zero:

$$\frac{\partial \mathcal{M}}{\partial \mathbb{E}_\theta[S \mid C = 1]} = 0 \quad (5)$$

This zero-gradient property mathematically formalizes the "Safety Illusion," leading to two catastrophic optimization failures in the saturation regime:

- **Masked Tail Risk:** A policy generating 4.2% fatal, high-speed crashes ($\mathbb{E}_\theta[S \mid C = 1] \gg 0$) achieves a near-perfect realism score, as the metric is entirely blind to the lethal consequences.

- **The Safety Penalty:** A genuinely safe policy that actively avoids all collisions ($P_\theta(C = 1) \to 0$) is mathematically penalized for distributional mismatch against the 4.2% dataset noise floor.

Consequently, a binary likelihood objective provides no informative gradients for safety alignment, necessitating a shift to a continuous severity metric.

4. Methodology

To bridge the resolution gap identified in binary metrics, we introduce a physically grounded evaluation framework. Rather than collapsing multi-agent interactions into a rollout-level boolean, we formulate a continuous, dimensionless severity score that evaluates the physical consequence of each pairwise collision event.

4.1. Event-Level Kinematic Abstraction

Instead of computing a single global collision flag for an evaluated agent, we extract pair-wise collision events. For any evaluated agent interacting with another object, we record a physical tuple $(v_{\text{rel}}, d, \Delta t)$, where v_{rel} is the relative impact velocity, d is the maximum penetration depth, and Δt is the continuous contact duration.

4.2. Collision Taxonomy and Noise Filtering

Not all geometric contacts represent meaningful safety failures. To prevent annotation ambiguity, bounding-box jitter, or pedestrian-dominant artifacts from skewing our severity analysis, we apply a lightweight noise filter immediately after event extraction.

Specifically, we classify an event as *Noise* if it is (i) a pedestrian–pedestrian contact, or (ii) a pedestrian–vehicle contact where the pedestrian's impact speed exceeds the vehicle's (e.g., a pedestrian running into a static or slower vehicle). All remaining events are retained for analysis. Unless otherwise specified, all subsequent references to *collisions*, *collision rates*, and *severity statistics* refer exclusively to this filtered set of meaningful collisions.

4.3. Structural-Damage-Dominant Severity (S)

For each retained collision event after Noise filtering, we define the Composite Severity S as a non-linear combination of the kinematic tuple. The metric is explicitly designed to heavily penalize structural damage and prolonged sticking, while remaining robust to the inherent noise and "teleportation" artifacts common in autoregressive sim-agents[33, 44, 50, 52].

$$S = m(v_{\text{rel}}) \cdot \delta(d) \cdot g(\Delta t) \tag{6}$$

The components are defined as follows:

1. Momentum Multiplier $m(v_{\text{rel}})$: The relative impact velocity is $v_{\text{rel}} = \|\vec{v}_1 - \vec{v}_2\|$. While empirical severity metrics like Δv and kinetic energy dissipation scale quadratically [21, 42, 43], such formulations easily amplify simulation instabilities (e.g., coordinate anomalies causing exploding values). Thus, we deliberately apply a bounded linear momentum scaling:

$$m(v_{\text{rel}}) = \frac{\min(\max(v_{\text{rel}}, v_{\min}), v_{\max})}{v_{\text{ref}}} \tag{7}$$

where v_{ref} is a normalization constant. The lower bound $v_{\min}$ ensures low-speed grazing is strictly penalized, while the upper bound $v_{\max}$ truncates unphysical "teleportation" artifacts without skewing the overall risk distribution.

2. Structural Damage Penalty $\delta(d)$: To distinguish catastrophic T-bone collisions from minor boundary grazing, we apply a squared penalty to the penetration depth d. This aligns with classic vehicle safety engineering where absorbed energy is approximately proportional to the square of the crush depth [5, 43]. This explicitly separates severe structural damage from minor boundary grazing:

$$\delta(d) = \left(\frac{\max(d - \epsilon, 0)}{d_{\text{ref}}} \right)^2 \tag{8}$$

where d_{ref} normalizes the depth, and ϵ is a minimal tolerance to absorb bounding-box discretization errors. The quadratic nature ensures that the depth d remains the dominant risk factor.

3. Duration Gating $g(\Delta t)$: Autoregressive sim-agents simulation and data collection often exhibit transient bounding-box jitter or "flickering" collisions [15, 33, 44]. To prevent a single-frame intersection from being weighted equally to a prolonged, multi-second failure (e.g., two vehicles permanently stuck together), we introduce a non-linear temporal filter with an absolute deadzone:

$$g(\Delta t) = \begin{cases} 0 & \Delta t \le t_{\text{res}} \\ \left(\frac{\Delta t - t_{\text{res}}}{t_{\text{noise}} - t_{\text{res}}} \right)^2 & t_{\text{res}} < \Delta t \le t_{\text{noise}} \\ 1 & \Delta t > t_{\text{noise}} \end{cases} \tag{9}$$

where t_{res} represents the minimum physical resolution of the simulation. Collisions lasting less than t_{res} are treated as pure computational noise and strictly zeroed out. Contacts lasting between t_{res} and t_{noise} undergo a smooth parabolic transition to avoid harsh step-changes in the severity manifold, while sustained contacts exceeding t_{noise} receive the full physical penalty.

By default, we set $v_{\text{ref}} = 5.0\,\text{m/s}$, $d_{\text{ref}} = 0.5\,\text{m}$, $v_{\min} = 1.0\,\text{m/s}$, $v_{\max} = 40.0\,\text{m/s}$, $t_{\text{res}} = 0.1\,\text{s}$, $t_{\text{noise}} = 0.2\,\text{s}$, and $\epsilon = 10^{-4}$.

4.4. Tail Risk Measurement

To prioritize the evaluation of catastrophic failures over minor artifacts, we employ the **Conditional Value-at-Risk**

Table 1. **Model Performance Summary:** Benchmark Likelihood and Severity-Oriented Safety Metrics. ↑ (↓) indicates that higher (lower) values are preferred. The best results are **bolded**, and the second best (excluding Log Replay) are underlined.

Model	WOSAC Likelihood Metrics				Collision & Severity Metrics (ours)		
	Realism Meta	Coll. Ind. Likelihood	Sim. Coll. Rate (%)	TTC Likelihood	Collision Rate (%)	Cond. CVaR_{95} $(S \mid C=1)$	CCM (Uncond. CVaR_{95})
Log Replay (GT)[15]	—	—	—	—	0.72	9.130	0.028
SMART[44]	0.7638	0.9643	4.24	0.8356	2.36	58.469	5.557
CAT-K[50]	**0.7657**	**0.9672**	**4.15**	**0.8362**	1.62	37.379	2.792
Safe CAT-K (ours)	0.7623	0.9637	4.42	0.8346	**1.51**	**19.138**	**1.611**

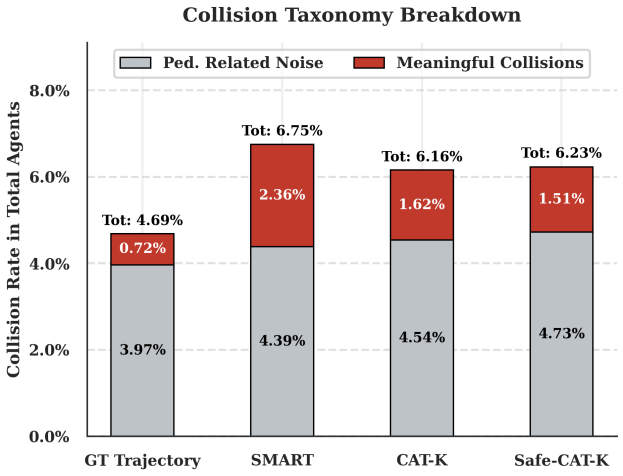

Figure 2. Breakdown of collision events by category. Stacked bars report category-specific event counts normalized by the evaluated-agent population, split into *Noise* (pedestrian-related noise events) and *Meaningful Collisions*. Numbers inside bars indicate the normalized rate of each category.

(CVaR)[35] at the $\alpha = 0.95$ level. While traditionally a financial metric, CVaR is increasingly recognized as a principled risk measure for autonomous systems, overcoming the "risk-neutral" blind spots of expected-value metrics [9]. For the severity distribution S conditioned on a meaningful collision ($C = 1$), we first define the Value-at-Risk (VaR) as the α-quantile:

$$\text{VaR}_\alpha = \inf\{s \in \mathbb{R} : P(S \leq s \mid C = 1) \geq \alpha\} \quad (10)$$

The CVaR_α is then calculated as the expected severity of the $(1-\alpha)$ worst-case events:

$$\text{CVaR}_\alpha(S \mid C = 1) = \mathbb{E}[S \mid S \geq \text{VaR}_\alpha, C = 1] \quad (11)$$

By setting $\alpha = 0.95$, this metric specifically targets the "long-tail" risk of high-impact collisions, effectively capturing severe physical consequences that are often obscured by mean-based statistics.

4.5. Composite Collision Metric

Assessing overall system safety requires capturing both accident frequency and magnitude. By assigning zero severity to collision-free instances and composite severity S to collision-positive ones, we form a global severity distribution. Its Unconditional Value-at-Risk (VaR) at level α is:

$$\text{VaR}_\alpha(S) = \inf\{s \in \mathbb{R} : P(S \leq s) \geq \alpha\} \quad (12)$$

We define our **Composite Collision Metric (CCM)** as the Unconditional CVaR, which captures the expected severity of the $(1-\alpha)$ worst-case events:

$$\text{CVaR}_\alpha(S) = \mathbb{E}[S \mid S \geq \text{VaR}_\alpha(S)] \quad (13)$$

A global mean severity ($\mathbb{E}[S]$) dangerously dilutes catastrophic collisions among overwhelmingly safe rollouts [27, 47]. To prevent this, CVaR_{95} ($\alpha = 0.95$) strictly evaluates tail-end performance. As a coherent risk measure, CVaR reliably captures extreme risks without the mathematical blind spots of standard VaR [2].

5. Experiments

This section evaluates whether binary collision likelihood remains a reliable safety signal in the high-score regime of closed-loop traffic simulation, and whether our severity-based analysis provides a more discriminative and optimization-relevant alternative. To this end, we organize the experiments into three stages. We first expose the failure mode of benchmark-aligned binary metrics under near-saturated WOSAC performance. We then test whether our metric can detect meaningful safety improvements that are nearly invisible to the benchmark. Finally, we analyze the long-tail risk profiles of different simulators to show why severity-aware evaluation is necessary for iteration-ready safety alignment.

5.1. Experimental Setup

Benchmark and evaluation protocol. All experiments are conducted on the WOMD[15] under the WOSAC evaluation protocol[31]. We follow the standard closed-loop rollout setting and report both benchmark-native metrics and

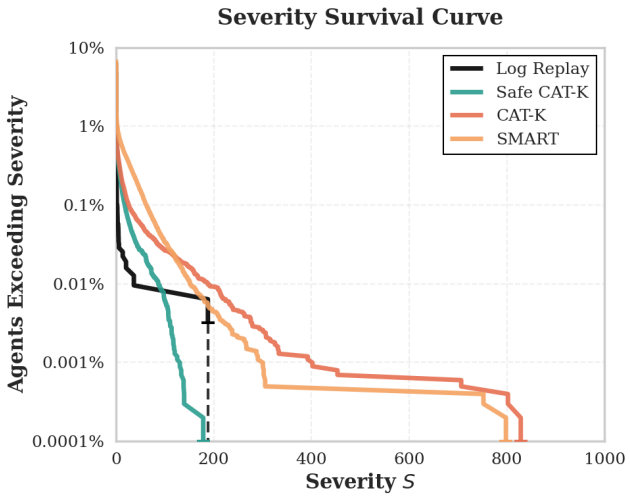

Figure 3. Survival function $P(S > s)$ of the composite collision severity. While models achieve similar binary collision likelihoods, their survival curves diverge significantly in the high-severity regime ($S \gg 0$), exposing hidden tail risks that are not evident from standard benchmark summaries.

our proposed severity-based metrics. Since our goal is not to replace realism evaluation but to examine its safety resolution, we always interpret the results jointly from two perspectives: benchmark-aligned realism and physically grounded collision consequence.

Implementation details. We use the default parameters and configurations to maintain consistency with the WOSAC leaderboard[31] evaluations. For the rollout generation, we generate 32 simulated rollouts per scenario, utilizing top-K sampling with $K = 32$. We fix the inference time sampling temperature to 1.0, others exactly following the original CAT-K setup[50]. We follow the evaluation setting in prior work [50] and perform severity analysis on a 2% subset of the WOMD[15] validation set, corresponding to 880 scenarios.

For the baseline models, we employ a 7M-parameter SMART model (SMART-tiny) trained via Behavior Cloning [44]. We reproduce the SMART-tiny weights locally using 8 RTX 5090 GPUs. To align with the training configuration in CAT-K, we use a per-GPU batch size of 2 combined with gradient accumulation to achieve an effective total batch size of 80[50]; all other hyperparameter settings remain identical to the original implementation. For the CAT-K model, we directly utilize the official pre-trained weights published by the authors[50].

Compared simulators. We consider four representative rollout sources. **Log Replay** [15]serves as a reference upper bound for distributional faithfulness, since it directly replays trajectories from the dataset. **SMART**[44] and **CAT-K**[50] are strong learned sim-agent baselines that achieve highly competitive WOSAC leaderboard performance and therefore represent the current high-score regime in which binary safety metrics are close to saturation. In addition, we construct **Safe CAT-K**, a safety-enhanced variant of CAT-K[50]. By incorporating dynamics-aware rejection sampling during rollout generation, we demonstrate a standard approach to improving model safety—specifically by mitigating severe and physically implausible collisions while maintaining the original driving behavior. Since this variant is constructed explicitly as a controlled intervention to validate our proposed metrics, rather than as a standalone architectural contribution, further investigation into this rejection strategy remains outside the scope of this paper and is therefore not discussed further.

Metric groups. We report three groups of metrics.

(1) WOSAC metrics.[31] These include the benchmark score and collision likelihood, which reflect how closely the rollout-level binary collision statistics match the ground-truth data distribution.

(2) Collision-frequency metrics. We report the collision rate after event-level noise filtering, which characterizes how often retained collision outcomes occur.

(3) Severity and Composite metrics. We report the conditional CVaR_{95} on the collision subset to measure pure collision severity. Furthermore, we report our CCM, defined as the unconditional CVaR_{95} over the full evaluated population, which comprehensively captures both collision frequency and severity.

Event extraction and aggregation. As defined in Sec. 4.2, we extract pairwise collision events and convert each event into a severity score using the tuple $(v_{\mathrm{rel}}, d, \Delta t)$.To reduce the influence of dataset artifacts and social crowding, we exclude pedestrian–pedestrian contacts and a subset of pedestrian–vehicle interactions in which the pedestrian actively runs into a static or slower-moving vehicle. Vehicle–vehicle collisions and the remaining semantically meaningful interactions are retained for analysis. Severity is first computed at the collision-event level and then aggregated over the agent population for reporting.

Evaluation objective. The experiments are designed to answer three questions. First, do benchmark-aligned binary metrics remain informative when several simulators already achieve similarly high WOSAC scores? Second, can our severity-based metrics respond to meaningful safety improvements that are nearly invisible to binary collision likelihood? Third, can tail-risk analysis reveal structural differences in failure behavior that are hidden by scalar benchmark scores?

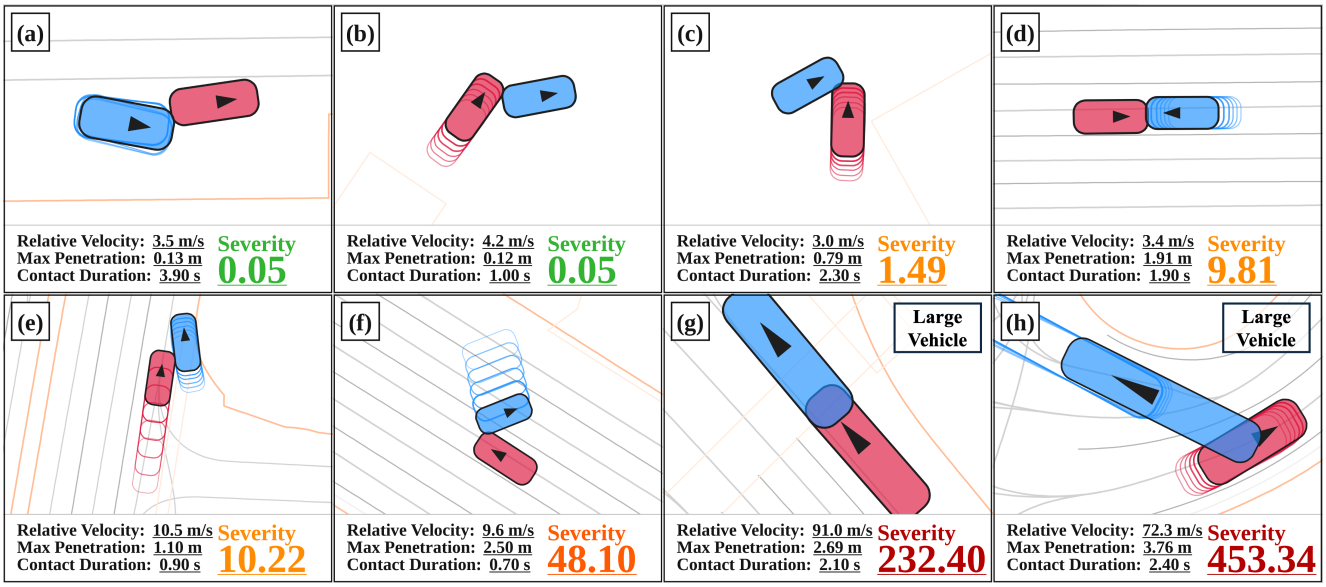

Figure 4. Qualitative visualization of collision severity. The events (a-h) are sampled from the rollouts of SMART, CAT-K, and Safe CAT-K, ordered from lowest to highest severity. In each subplot, the red and blue boxes represent the colliding vehicles, with arrows indicating their current headings. The fading transparent footprints trace the vehicles' trajectories over the past 0.5 seconds (at 10Hz) to illustrate pre-collision kinematics. While (a) and (b) are near-boundary and low-severity contacts, (f), (g), and (h) expose severe, physically implausible collisions caused by sampling unphysical tokens in Next-Token Prediction (NTP) models. The background is the map feature in the scenario.

5.2. Experiment Results

Near-saturated benchmarks. We first examine whether models with similar benchmark-facing performance remain distinguishable under severity-aware evaluation. As shown in Fig. 2 and Table 1, several models achieve competitive collision-likelihood and realism scores under the benchmark protocol. For instance, SMART[44] and CAT-K[50] achieve high WOSAC collision likelihoods (0.9643 and 0.9672, respectively) alongside comparable realism meta-scores (0.7638 and 0.7657). It is worth noting that the WOSAC collision rates differ from those in Fig. 2 because WOSAC evaluates only a pre-selected subset of agents of interest rather than all simulated agents. While this approach filters out certain noise, it severely restricts the metric's scope.

However, this benchmark-level similarity breaks down once binary collision events are decomposed into taxonomy-filtered outcomes and continuous severity statistics. While the Log Replay (GT) maintains a very low conditional CVaR_{95} of 9.130, SMART[44] and CAT-K[50] reach 58.469 and 37.379, respectively. This reveals that despite scoring similarly on the benchmark, they generate fundamentally different, and far more severe, physical consequences.

Safety intervention. We next test whether the proposed metrics are sensitive to deliberate safety-oriented improvement. To this end, we introduce Safe CAT-K, a controlled intervention on CAT-K that suppresses severe collisions while preserving the overall rollout distribution as much as possible. This provides a direct test of whether the evaluation protocol responds when a simulator becomes physically safer.

As summarized in Table 1, CAT-K[50] and Safe CAT-K show negligible differences under WOSAC-style metrics, with collision likelihood shifting only slightly from 0.9672 to 0.9637, and the realism meta-score dropping marginally to 0.7623. Under severity-aware evaluation, however, the improvement becomes massive. Safe CAT-K reduces the conditional CVaR_{95} by nearly half (from 37.379 to 19.138) and drastically lowers the CCM from 2.792 to 1.611. Together with the collision composition shown in Fig. 2, these results indicate that benchmark-level realism can remain largely unchanged even when the physical consequence of collisions is meaningfully reduced.

Long-tail risk. We further study whether severity-aware evaluation better exposes rare but catastrophic failures. To this end, we analyze the tail behavior of collision severity distributions through the survival function of severity exceedance. This perspective complements the aggregate statistics in Table 1 by directly characterizing the persis-

tence of high-consequence events.

The resulting curves in Fig. 3 reveal substantial separation in the high-severity regime. The replayed reference decays rapidly, indicating relatively light tails, whereas SMART[44] and CAT-K[50] retain much heavier tails, implying a larger probability of severe collisions. Safe CAT-K lies between CAT-K[50] and the reference, reflecting partial mitigation of catastrophic failures. This separation is much less explicit under binary collision summaries alone, but becomes visually and quantitatively clearer once the analysis focuses on the tail of the severity distribution.

5.3. Qualitative Analysis of Collision Severity

To relate the quantitative metrics to observable simulation behavior, we present a qualitative analysis of collision events. Fig. 4 shows eight representative vehicle-to-vehicle collisions sampled from the rollouts of SMART, CAT-K, and Safe CAT-K, ordered by increasing severity score S. The examples illustrate how the continuous metric separates low-impact contacts from higher-consequence collision events. We note that the visualization captures the first frame at which a collision is detected, rather than the frame of maximum penetration.

As shown in Fig. 4(a) and (b), the lowest severity scores ($S \approx 0.05$) correspond to slight scrapes or small bounding-box overlaps. These cases are visually consistent with minor contact events near the boundary of collision detection. Figs. 4(c), (d), and (e) ($S \in [1.0, 20.0]$) show more evident vehicle-to-vehicle interactions, including rear-end and intersection-style collisions with moderate relative motion and penetration.

At higher severity levels, the visual patterns change noticeably. Fig. 4(f) ($S \approx 48.1$) already shows a larger relative impact and a less smooth pre-collision motion pattern than the lower-severity cases. Figs. 4(g) and (h) ($S \in [200, 500]$) exhibit abrupt motion changes immediately before impact, together with very large overlap and substantially higher instantaneous speed. In these cases, the colliding vehicles reach speeds above $70\,\mathrm{m/s}$, and the resulting penetration is much larger than in the lower-severity examples.

Overall, the examples in Fig. 4 show a consistent progression in visible collision consequence as S increases: from slight contact, to moderate interaction, to large-overlap, high-speed events. This qualitative trend is consistent with the quantitative severity statistics in Table 1 and the tail behavior shown in Fig. 3.

6. Discussion

Main findings. Our findings directly answer the three questions posed in Sec. 5. First, in the high-score regime of modern generative sim-agents, benchmark-aligned realism does not equate to physical safety; binary metrics create a "safety illusion" by masking severe long-tail risks. Second, the Safe CAT-K intervention shows that severity-aware metrics can capture meaningful safety improvements that remain nearly invisible under standard benchmark indicators [31]. Third, survival analysis reveals structural differences in failure behavior that scalar collision rates completely obscure [44, 50].

Broader implication. As autonomous driving increasingly depends on high-fidelity, long-horizon simulation, the definition of a "good" simulator must evolve beyond matching dataset-level statistics alone [41, 47]. In this setting, a physically grounded and highly discriminative collision metric is essential, both for rigorous benchmark evaluation and for simulator refinement [32, 37].

Future direction. An important next step is to move severity-aware metrics from passive evaluation toward active safety alignment. While unconditional CVaR provides a robust tail-sensitive evaluation signal [2, 35], its underlying kinematic components could be further restructured into a joint risk objective for reinforcement learning, helping prevent policies from trading catastrophic failures for more frequent minor collisions [9]. Specifically, we aim to model missing crash types and pre-crash human behaviors, followed by an expansion into near-miss risks. This progression is essential for providing a continuous safety signal that captures critical semantics currently invisible to binary metrics [18, 22].

7. Conclusion

By analyzing the fundamental limitations of binary collision metrics, we reveal how they create a safety illusion that obscures critical risks in high-fidelity autonomous driving simulations. To address this limitation, we decompose collision events into interpretable kinematic components and aggregate them into continuous severity-based statistics, including the proposed CCM.

Experiments on WOSAC reveal three main findings. First, models with similar benchmark-facing realism can still exhibit sharply different physical safety profiles. Second, meaningful safety improvements, such as those induced by Safe CAT-K, are clearly captured by severity-aware metrics but remain largely invisible to standard binary indicators. Third, tail-risk and survival analysis expose structural differences in failure behavior that scalar collision rates cannot represent.

As autonomous driving increasingly relies on simulation for both testing and iterative model improvement, evaluation must move beyond binary formulations toward metrics that preserve physical consequence and long-tail risk, which serves as a critical next step for accurately modeling safe driving behaviors.

Acknowledgments

The work presented in this paper was jointly supported by the Faculty Interdisciplinary Fund of the University of Hong Kong (HKU) and the HKU–USI Fellowship Grant from the Urban Systems Institute (USI).

References

[1] Ehsan Ahmadi and Hunter Schofield. Rlftsim: Multi-agent traffic simulation via reinforcement learning fine-tuning (technical report for waymo open sim agents challenge). 2025. 3

[2] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999. 5, 8

[3] Jonathan Booher, Khashayar Rohanimanesh, Junhong Xu, Vladislav Isenbaev, Ashwin Balakrishna, Ishan Gupta, Wei Liu, and Aleksandr Petiushko. Cimrl: Combining imitation and reinforcement learning for safe autonomous driving. *arXiv preprint arXiv:2406.08878*, 2024. 2

[4] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 2

[5] Kenneth L Campbell. Energy basis for collision severity. Technical report, SAE Technical Paper, 1974. 4

[6] Wei-Jer Chang, Akshay Rangesh, Kevin Joseph, Matthew Strong, Masayoshi Tomizuka, Yihan Hu, and Wei Zhan. Spacer: Self-play anchoring with centralized reference models. *arXiv preprint arXiv:2510.18060*, 2025. 2, 3

[7] Di Chen, Meixin Zhu, Hao Yang, Xuesong Wang, and Yinhai Wang. Data-driven traffic simulation: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, 9(4):4730–4748, 2024. 2

[8] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022. 2

[9] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015. 5, 8

[10] PJ Cooper. Experience with traffic conflicts in canada with emphasis on "post encroachment time" techniques. In *International calibration study of traffic conflict techniques*, pages 75–96. Springer, 1984. 2

[11] Daphne Cornelisse. Human-likeness metrics for autonomous agents: are we measuring the right thing? Substack, 2025. Blog post analyzing the Waymo Open Sim Agent Challenge (WOSAC) realism benchmark. 2

[12] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024. 2

[13] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2, 3

[14] Christer Ericson. *Real-time collision detection*. Crc Press, 2004. 12

[15] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9710–9719, 2021. 2, 3, 4, 5, 6, 13, 14

[16] Shuo Feng, Xintao Yan, Haowei Sun, Yiheng Feng, and Henry X Liu. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature communications*, 12(1):748, 2021. 2

[17] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953):620–627, 2023. 2

[18] Douglas Gettman, Lili Pu, Tarek Sayed, Steven G Shelby, et al. Surrogate safety assessment model and validation. Technical report, Turner-Fairbank Highway Research Center, 2008. 2, 8

[19] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36:7730–7742, 2023. 2, 3

[20] Tao Han, Wanghan Xu, Junchao Gong, Xiaoyu Yue, Song Guo, Luping Zhou, and Lei Bai. Infgen: A resolution-agnostic paradigm for scalable image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17941–17950, 2025. 2

[21] Hans Hauschild, Dale Halloway, and Frank Pintar. Delta-v slope as an indicator of injury. *Traffic injury prevention*, 22 (sup1):S165–S169, 2021. 4

[22] John C Hayward. Near miss determination through use of a scale of danger. 1972. 8

[23] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023. 2

[24] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7953–7963, 2023. 2

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

vances in Neural Information Processing Systems, 37:819–844, 2024. 2

[26] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 2

[27] Nidhi Kalra and Susan M Paddock. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation research part A: policy and practice*, 94:182–193, 2016. 5

[28] Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic goodhart: regularizing rlhf with kl divergence does not mitigate heavy-tailed reward misspecification. *Advances in Neural Information Processing Systems*, 37:14608–14633, 2024. 2

[29] Henry Liu, Zhong Cao, Xintao Yan, Shuo Feng, and Qiujing Lu. Autonomous vehicles: A critical review (2004-2024) and a vision for the future. *Authorea Preprints*, 2025. 2

[30] Henry X Liu, Xintao Yan, Haowei Sun, Tinghan Wang, Zhijie Qiao, Haojie Zhu, Shengyin Shen, Shuo Feng, Greg Stevens, and Greg McGuire. Behavioral safety assessment towards large-scale deployment of autonomous vehicles. *arXiv preprint arXiv:2505.16214*, 2025. 2

[31] Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, et al. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36:59151–59171, 2023. 2, 3, 5, 6, 8, 12, 14, 15

[32] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022. 2, 8

[33] Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. *arXiv preprint arXiv:2312.04535*, 2023. 3, 4

[34] Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A Koepke, Zeynep Akata, and Andreas Geiger. Plant: Explainable planning transformers via object-level representations. *arXiv preprint arXiv:2210.14222*, 2022. 2

[35] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000. 5, 8

[36] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. 2

[37] Hunter Schofield, Mohammed Elmahgiubi, Kasra Rezaee, and Jinjun Shan. Beyond simulation: Benchmarking world models for planning and causality in autonomous driving. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1308–1314. IEEE, 2025. 2, 8

[38] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022. 3

[39] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3955–3971, 2024. 3

[40] Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10400–10409, 2021. 3

[41] Shuhan Tan, John Lambert, Hong Jeon, Sakshum Kulshrestha, Yijing Bai, Jing Luo, Dragomir Anguelov, Mingxing Tan, and Chiyu Max Jiang. Scenediffuser++: City-scale traffic simulation via a generative world model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1570–1580, 2025. 2, 8

[42] Ada H Tsoi and Hampton C Gabler. Evaluation of vehicle-based crash severity metrics. *Traffic injury prevention*, 16 (sup2):S132–S139, 2015. 4

[43] Dario Vangi. Simplified method for evaluating energy loss in vehicle collisions. *Accident Analysis & Prevention*, 41(3): 633–641, 2009. 4

[44] Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time motion generation via next-token prediction. *Advances in Neural Information Processing Systems*, 37:114048–114071, 2024. 2, 3, 4, 5, 6, 7, 8, 13, 14

[45] Lingyu Xiao, Jiang-Jiang Liu, Xiaoqing Ye, Wankou Yang, and Jingdong Wang. Easychauffeur: A baseline advancing simplicity and efficiency on waymax. *arXiv preprint arXiv:2408.16375*, 2024. 3

[46] Lingyu Xiao, Jiang-Jiang Liu, Sen Yang, Xiaofan Li, Xiaoqing Ye, Wankou Yang, and Jingdong Wang. Learning multiple probabilistic decisions from latent world model in autonomous driving. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1279–1285. IEEE, 2025. 2, 3

[47] Xintao Yan, Zhengxia Zou, Shuo Feng, Haojie Zhu, Haowei Sun, and Henry X Liu. Learning naturalistic driving environment with statistical realism. *Nature communications*, 14 (1):2037, 2023. 2, 5, 8

[48] Xintao Yan, Shuo Feng, Haowei Sun, and Henry X Liu. Distributionally consistent simulation of naturalistic driving environment for autonomous vehicle testing. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 2

[49] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15222–15232, 2021. 2

[50] Zhejun Zhang, Peter Karkus, Maximilian Igl, Wenhao Ding, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Closed-loop supervised fine-tuning of tokenized traffic models. In

Proceedings of the Computer Vision and Pattern Recognition Conference, pages 5422–5432, 2025. 2, 3, 4, 5, 6, 7, 8, 13, 14, 15

[51] Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[52] Zikang Zhou, Zihao Wen, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Qcnext: A next-generation framework for joint multi-agent trajectory prediction. *arXiv preprint arXiv:2306.10508*, 2023. 4

[53] Zikang Zhou, HU Haibo, Xinhong Chen, Jianping Wang, Nan Guan, Kui Wu, Yung-Hui Li, Yu-Kai Huang, and Chun Jason Xue. Behaviorgpt: Smart agent simulation for autonomous driving with next-patch prediction. *Advances in Neural Information Processing Systems*, 37:79597–79617, 2024. 2, 3

Beyond Binary Metrics: Unveiling the Safety Illusion in Autonomous Driving Simulation

Supplementary Material

A. Collision Event Detection Details

Penetration Depth Calculation (SAT Approximation) In the context of the Separating Axis Theorem (SAT)[14], the physical penetration depth d is geometrically defined as the magnitude of the Minimum Translation Vector (MTV)—the shortest distance required to separate two colliding convex shapes.

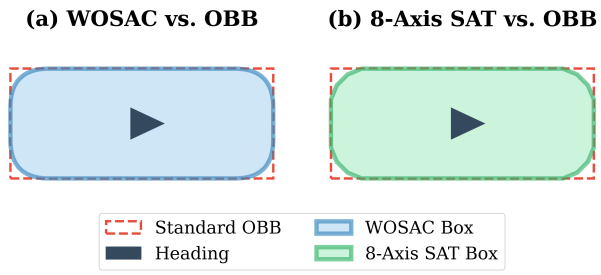

Figure S1. Comparison of bounding box geometries used for collision detection. The red dashed line represents the Standard Oriented Bounding Box (OBB). **(a)** The WOSAC[31] protocol utilizes an exact continuous rounding approach (blue). **(b)** Our proposed method uses an 8-axis SAT[14] approximation (green), which forms a 16-sided polygon to efficiently mimic the rounded geometry without the computational overhead of non-linear intersection checks.

For an evaluated agent pair (i, j) at a collision timestep, we calculate the overlap along each normalized test axis $\mathbf{a} \in \mathcal{A}_{16}$. Let $\mathbf{p}_i$ and $\mathbf{p}_j$ be the centroid coordinates, and $\mathcal{B}_{core}$ denote the shrunk core rectangle with half-extents $\mathbf{e} = [l_{core}/2, w_{core}/2]^T$. The projected overlap on a single axis $\mathbf{a}$ is given by:

$$\text{Overlap}(\mathbf{a}) = \rho_i(\mathbf{a}) + \rho_j(\mathbf{a}) + 2r - |(\mathbf{p}_j - \mathbf{p}_i) \cdot \mathbf{a}| \quad \text{(S1)}$$

where $2r$ is the combined corner-rounding inflation factor (with $r = 0.7$m), and $\rho(\mathbf{a})$ is the projected radius of the core rectangle onto the test axis, calculated as:

$$\rho(\mathbf{a}) = e_x|\mathbf{u}_x \cdot \mathbf{a}| + e_y|\mathbf{u}_y \cdot \mathbf{a}| \quad \text{(S2)}$$

where $\mathbf{u}_x$ and $\mathbf{u}_y$ are the local orthogonal heading vectors of the agent.

A collision is confirmed if $\text{Overlap}(\mathbf{a}) > 0$ for all 16 axes. The final penetration depth d is then extracted as the minimum overlap across all tested directions, representing the path of least resistance to separate the vehicles:

$$d = \min_{\mathbf{a} \in \mathcal{A}_{16}} \text{Overlap}(\mathbf{a}) \quad \text{(S3)}$$

Kinematic Extraction and Temporal Aggregation While the penetration depth d captures the maximum spatial severity during an event, our composite metric also depends on the relative impact velocity v_{rel} and the contact duration Δt.

In autoregressive rollouts, vehicles may exhibit unphysical behaviors (e.g., "teleportation" or abrupt kinematic jumps) after an initial crash. To ensure kinematic stability and isolate the true cause of the accident, the relative velocity is strictly sampled at the *first frame* of contact t_{start}:

$$v_{rel} = ||\mathbf{v}_i(t_{start}) - \mathbf{v}_j(t_{start})||_2 \quad \text{(S4)}$$

The continuous contact duration is simply accumulated over the contiguous frames where the SAT overlap remains positive: $\Delta t = (t_{end} - t_{start} + 1) \cdot \Delta t_{sim}$, where $\Delta t_{sim} = 0.1$s is the simulation resolution.

Taxonomy-Aware Noise Filtering To prevent dataset artifacts and harmless social crowding from skewing the severity distribution, we apply a semantic noise filter after the event extraction. A collision tuple is classified as noise and excluded from the final evaluation if it meets any of the following criteria:

1. **Pedestrian-Pedestrian Interaction:** Both agents are pedestrians, which predominantly stems from bounding-box jitter or dense crowd approximations rather than severe safety failures.
2. **Active Pedestrian Hit:** The collision involves a pedestrian and a vehicle, but the pedestrian's absolute speed is greater than or equal to the vehicle's speed at the moment of impact ($||\mathbf{v}_{ped}(t_{start})||_2 \geq ||\mathbf{v}_{veh}(t_{start})||_2$). This filters out scenarios where a pedestrian actively runs into a static or slower-moving ego vehicle.

B. Metric Behavior and Hyperparameter Robustness

To examine whether the proposed severity formulation yields a stable and physically interpretable evaluation signal, we explicitly analyze the marginal behavior of the Composite Severity S and its sensitivity to hyperparameter selections.

B.1. Monotonicity of the Composite Severity

A reliable safety metric must ensure that as the physical risk of a collision increases, the penalized severity strictly increases or remains flat (i.e., non-decreasing). We verify

Table S1. Sensitivity of CCM (Unconditional CVaR_{95}) evaluated across the full validation set under varying reference depth (d_{ref}) and velocity (v_{ref}) settings. The relative safety rankings remain identical.

Model	**Default**	Depth Variation ($v_{ref} = 5.0$)		Velocity Variation ($d_{ref} = 0.5$)	
	($d_{ref} = 0.5, v_{ref} = 5.0$)	$d_{ref} = 0.25$	$d_{ref} = 1.0$	$v_{ref} = 2.5$	$v_{ref} = 10.0$
Log Replay (GT)[15]	0.028	0.111	0.007	0.055	0.014
Safe CAT-K (ours)	**1.611**	**6.446**	**0.403**	**3.223**	**0.806**
CAT-K[50]	2.792	11.169	0.698	5.585	1.396
SMART[44]	5.557	22.230	1.389	11.115	2.779

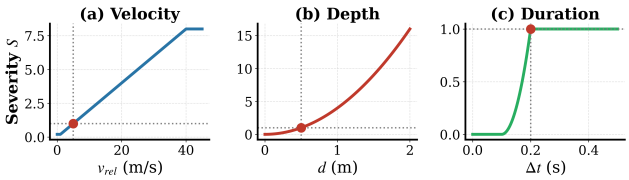

Figure S2. Marginal response curves of the **Composite Severity** S, with the default hyper-parameters defined in this manuscript. The **red dots** denote the normalization anchor points where the metric components evaluate to exactly 1.0.

this by isolating each kinematic component: relative velocity v_{rel}, penetration depth d, and contact duration Δt, then holding the other two constant at their reference values.

As illustrated in Figure S2, the severity score S exhibits strict monotonic behavior with respect to each risk factor:

- **Impact Velocity (v_{rel}):** The severity increases linearly with impact speed, bounded artificially at $v_{max} = 40.0$ m/s to prevent simulation coordinate anomalies ("teleportation" bugs) from exploding the metric gradient.
- **Penetration Depth (d):** The severity grows quadratically with d, ensuring that fatal structural penetrations (e.g., T-bone crashes) are penalized exponentially more than superficial boundary grazing.
- **Contact Duration (Δt):** The temporal filter effectively zeros out pure computational noise ($\Delta t \leq 0.1$ s), smoothly transitions via a parabolic curve to prevent harsh gradient steps, and flatlines at a multiplier of 1.0 for sustained contacts.

B.2. Choice of Reference Values

The reference parameters $v_{ref} = 5.0$ m/s and $d_{ref} = 0.5$ m are not arbitrarily chosen; they are deliberately designed to establish a normalized, semantic threshold for what constitutes a physically "meaningful" collision. Specifically, when a collision occurs at exactly the reference relative velocity and reference penetration depth, with a valid sustained duration ($\Delta t > t_{noise}$), each mathematical component evaluates to exactly 1.0. Consequently, the overall severity score is strictly $S = 1.0 \times 1.0 \times 1.0 = 1.0$.

This normalization gives $S = 1.0$ an intuitive interpretation as a reference-scale collision: values substantially above 1 indicate increasingly severe events, whereas values below 1 generally correspond to lighter contacts near the low-severity regime. By normalizing the formula around $S = 1.0$, the metric ensures that values numerically exceeding the reference are explicitly recognized as consequential safety violations.

B.3. Sensitivity to Hyperparameter Variations

To verify that the relative safety rankings of the evaluated simulators are robust to the exact choice of anchor hyperparameters, we re-evaluate the full validation set and recalculate the CCM (Unconditional CVaR_{95}) across varying reference depths ($d_{ref} \in \{0.25, 0.5, 1.0\}$) and reference velocities ($v_{ref} \in \{2.5, 5.0, 10.0\}$).

As shown in Table S1, while adjusting d_{ref} or v_{ref} changes the absolute scale of the CCM scores due to the altered normalization bounds, the relative ranking of the simulators (*Log Replay* >*Safe CAT-K* >*CAT-K* >*SMART*) remains strictly consistent across all configurations. The presence of non-linear clipping bounds in our formula (e.g., minimum velocity thresholds and spatial tolerances) introduces slight empirical variations compared to pure theoretical scaling, yet the monotonic separability of the models is perfectly preserved. This demonstrates that our conclusion regarding the "safety illusion" is mathematically robust, physically grounded, and not an artifact of hyperparameter tuning.

C. Robustness to Semantic Noise

This section investigates the impact of taxonomy-aware noise filtering on the evaluation landscape. By comparing safety metrics computed on the raw, unfiltered collision data against our semantically filtered subset, we demonstrate that while raw collision rates are highly sensitive to dataset artifacts, our proposed severity metrics (**Conditional CVaR_{95} and CCM**) possess an intrinsic mathematical immunity to such noise.

Table S2. Model Performance Summary Evaluated **Without** Taxonomy-Aware Noise Filtering. While the *Raw Total Collision Rate* is significantly inflated by semantic noise (e.g., pedestrian jitter), the severity-focused metrics (**Cond. CVaR$_{95}$** and **CCM**) and their relative rankings remain **identical** to the filtered results in the main text. This demonstrates that our framework is mathematically robust against low-severity dataset artifacts.

Model	WOSAC Likelihood Metrics				Unfiltered Collision & Severity Metrics (ours)		
	Realism Meta↑	Coll. Ind. Likelihood ↑	Sim. Coll. Rate (%) ↓	TTC Likelihood↑	Raw Total Col. Rate (%)↓	Cond. CVaR$_{95}$ $(S \mid C=1)\downarrow$	CCM (Uncond. CVaR$_{95}$)↓
Log Replay (GT)[15]	—	—	—	—	4.69	9.130	0.028
SMART[44]	0.7638	0.9643	4.24	0.8356	6.75	58.469	5.557
CAT-K[50]	**0.7657**	**0.9672**	**4.15**	**0.8362**	**6.16**	37.379	2.792
Safe CAT-K (ours)	0.7623	0.9637	4.42	0.8346	6.23	**19.138**	**1.611**

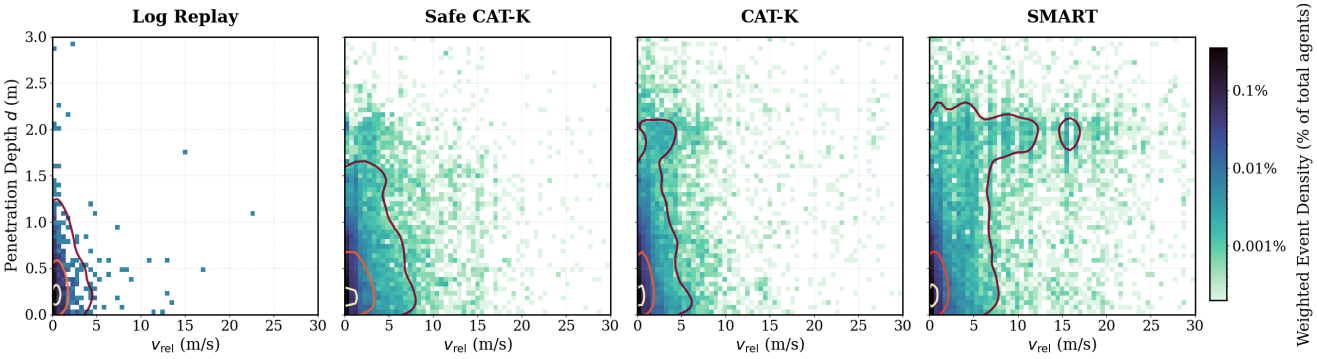

Figure S3. Visualization of collision severity via joint kinematic distributions. By plotting relative impact velocity (v_{rel}) against penetration depth (d), the heatmaps reveal the broad spectrum of physical collision behaviors—from minor contacts to extreme structural penetrations—across the evaluated models.

C.1. Mathematical Robustness of Severity Metrics

In the standard WOSAC[31] evaluation, all intersecting bounding boxes are treated as collisions. However, a significant portion of these instances constitutes *semantic noise*, primarily arising from pedestrian-related bounding box jitter or harmless crowding where no physical impact occurs. A key design objective for a reliable safety metric is the ability to distinguish these benign artifacts from catastrophic failures.

As shown in Table S2, our severity-based metrics (**Cond. CVaR$_{95}$** and **CCM**) are **empirically robust to weak semantic noise**. Despite the raw total collision rates inflating significantly due to the inclusion of pedestrian artifacts, the values for Cond. CVaR$_{95}$ and CCM remain **identical** to those reported in the filtered results of the main text.

This robustness is a direct consequence of the Conditional Value-at-Risk formulation. Since the severity S of pedestrian-related jitter or minor crowding is near zero ($S \approx 0$), these events consistently populate the lower 95% of the global severity distribution. Because CVaR$_{95}$ specifically computes the expectation of the top 5% tail risk, these low-severity noise points are mathematically excluded from the calculation. Consequently, our framework acts as an inherent, severity-aware filter, ensuring that safety rankings are driven exclusively by severe, meaningful accidents rather than dataset-specific labeling artifacts.

D. Joint Kinematic Distributions

D.1. Exposing Model-Specific Failure Modes

To further elucidate the Resolution Gap inherent in binary collision metrics, Figure S3 visualizes the joint kinematic distribution of relative impact velocity (v_{rel}) and penetration depth (d). Event densities are weighted by the total simulated agent population to reflect true probability mass. Consistent with the quantitative findings in our main text, while generative sim-agents achieve nearly indistinguishable benchmark-facing collision likelihoods, their joint kinematic profiles reveal fundamentally divergent physical behaviors. The Ground Truth (Log Replay) serves as an empirical noise floor, with collisions strictly localized near the origin ($v_{rel} < 5$ m/s, $d < 0.5$ m). In contrast, models like SMART [44] and CAT-K [50] exhibit severe long-tail risks, spreading significantly into high-severity domains with clusters of catastrophic, high-speed structural penetrations.

D.2. Capturing Semantic Shifts via Safety Interventions

The practical necessity of evaluating individual kinematic components becomes apparent when analyzing the Safe CAT-K intervention. Relative to CAT-K[50], Safe CAT-K exhibits visibly fewer events in the extreme high-speed / deep-penetration region, together with a comparatively denser concentration in the moderate-severity regime. This pattern is consistent with the quantitative reduction in tail severity reported in the main text.

D.3. Practical Reading of CCM

This visual evidence directly corroborates the core conclusion of our main text: benchmark-aligned binary metrics create a "safety illusion." Under standard WOSAC evaluation[31], trading a fatal 20 m/s T-bone crash for a 5 m/s minor scrape yields zero mathematical improvement, as both map identically to a binary indicator $C = 1$. Because Safe CAT-K simply shifts the mass of the distribution from the catastrophic tail into the moderate mid-range, its overall binary collision frequency remains largely static. However, by explicitly decoupling the evaluation into relative velocity and penetration depth, we demonstrate that the physical safety of the system has fundamentally improved. The proposed CCM successfully captures this semantic shift, visually and quantitatively proving that severity-aware analysis is a practically indispensable tool for genuine safety alignment.