Beyond Collision: A Human-Centric Process Safety Evaluation and Trajectory Supervision Framework for Autonomous Driving

Yunwei Li[1⋆], Ruilin Yu[1], Xiangyang Ye[1], and Siyu Wu[1]

School of Vehicle and Mobility, Tsinghua University, Beijing, China
li-yw23@mails.tsinghua.edu.cn

Abstract. Safety of the intended functionality (SOTIF) evaluations for autonomous driving systems often reduce acceptance to outcome-level evidence, such as whether a scenario is collision-free or whether a simple criticality threshold is crossed. We argue that this misses a central safety question: did the system complete the human-relevant driving process safely before the final outcome? This paper proposes a human-centric multi-stage evaluation framework for intersection driving. The framework decomposes a scenario into approach, in-intersection, and exit stages, and evaluates each stage through safe-driving-procedure completion, collision risk, behavioral feasibility, and efficiency. In a CARLA–Autoware software-in-the-loop case study, the framework exposes hidden process risk in a collision-free case, and diagnoses failure modes across 15 safety-critical scenarios where Autoware passes only 53.3% of cases. We further show that human-process signals can serve as a trajectory safety guardrail for end-to-end planning: on 504 OpenDriveVLA intersection samples, a human-process selector reduces collision steps from 19 to 12.

Keywords: Autonomous driving safety · SOTIF · Human-centric evaluation · Trajectory selection

1 Introduction

Autonomous driving systems (ADS) are increasingly evaluated in complex urban scenarios where functional failures are not the only source of danger. Functional safety addresses hazards caused by malfunctioning behavior [13]; under the safety of the intended functionality (SOTIF) view, risk may also arise from perception limits, algorithmic insufficiency, or environmental disturbances even when the system operates as designed [14]. This makes the acceptance criterion for safety testing a central question: what evidence is sufficient to conclude that an ADS behavior is safe enough?

Common acceptance criteria remain outcome-oriented. A test is often summarized by collision occurrence, pass rate, minimum time-to-collision (TTC), or related criticality scores [10, 34]. Such metrics are useful, but incomplete.

⋆ Corresponding author.

A vehicle may avoid collision only through late steering or harsh braking after missing an earlier perception or decision obligation. Conversely, a trajectory may match human ground truth poorly while being safer under interaction risk. In both cases, outcome-only criteria hide the intermediate process that should determine whether the system behaved defensively.

We therefore reframe SOTIF acceptance from an outcome-based question to a process-oriented one. Human drivers do not merely "avoid collision" at intersections; they execute a sequence of safe driving procedures (SDPs): observing lanes, traffic lights, and vulnerable road users; judging priority and occlusion; and controlling speed, acceleration, and steering smoothly. These procedures form a natural benchmark for white-box safety evaluation because they identify which subtask failed, when it failed, and whether the final outcome masks a latent risk.

This paper makes three contributions. First, we formulate a three-stage, four-dimension taxonomy for process-oriented intersection evaluation, covering approach, in-intersection, and exit stages, and scoring SDP completion, risk, behavioral feasibility, and efficiency. Second, we apply this taxonomy retrospectively to a full-stack Autoware [16] case study, revealing a collision-free but unsafe case, a multi-conflict failure case, and scenario-library weaknesses obscured by pass/fail testing. Third, we show that the human-centric signals can be used as a trajectory safety guardrail for planner: a human-process selector re-ranks multiple OpenDriveVLA candidates and reduces collision metrics without retraining the planner.

2 Related Work

ADS safety evaluation. Criticality metrics such as TTC and its variants provide compact indicators of imminent collision risk [10,21,32]. Broader taxonomies distinguish time-scale, distance-scale, acceleration-scale, and probabilistic measures, but most are still computed from trajectories or final scene states rather than from task completion evidence [34]. Scenario-based testing and SOTIF engineering additionally emphasize that unsafe behavior can arise without component faults, motivating tests that expose performance limitations and misuse cases [14, 18, 22, 31]. Simulation platforms such as CARLA, LGSVL/SVL, and SUMO enable scalable closed-loop testing in urban traffic [5, 19, 26]. However, even rich simulation is often summarized by collision, route completion, or aggregate discomfort, making it difficult to identify which perception, decision, or control obligation failed. Our framework keeps criticality metrics but embeds them in a staged process evaluation.

Human-centric driving models. Human-driver baselines are widely used in SOTIF-style acceptance criteria because they connect technical performance to socially meaningful driving competence. Classic driver models split behavior into strategic, tactical, and operational levels [23], and situation-awareness theory emphasizes perception, comprehension, and projection as prerequisites for safe action [6]. Car-following and lane-changing models further encode human-

like gap, comfort, and interaction constraints [9,17,30]. Existing ADS evaluations often use aggregate human crash or pass rates, which makes the human baseline an outcome comparator. We instead operationalize human driving as stage-specific perception, decision, and control obligations, allowing the evaluator to identify hidden process risks in otherwise successful trials.

Planning and trajectory selection. Full-stack modular systems such as Autoware [16] expose intermediate modules, while end-to-end planners increasingly optimize perception, prediction, occupancy, and planning jointly. Representative systems include ST-P3 [11], UniAD [12], VAD [15], TransFuser [25], TCP [35], and language-conditioned driving models such as DriveLM and OpenDriveVLA [28,36]. Large datasets and benchmarks, including KITTI, nuScenes, Waymo Open Dataset, Argoverse, and nuPlan, provide standardized perception and planning settings [2–4,8,29]. End-to-end systems can produce multiple plausible trajectories, but selecting the safer candidate requires structured criteria beyond imitation loss. Safety filters such as responsibility-sensitive safety and safety force fields are related in spirit [24, 27]; our offline study uses process-inspired costs to rank a fixed candidate set rather than claiming a deployable safety filter.

Interpretability and feature selection. The second application studies sensitivity to process-inspired signal groups in trajectory selection. We therefore draw on model reliance and permutation importance, which evaluate how prediction quality changes when features or feature groups are perturbed [1, 7]. SHAP-style additive explanations are another common option [20]; to accommodate different model architectures, we utilize random-forest permutation importance to evaluate feature sensitivity. We treat the resulting feature-group results as exploratory ablations on the evaluated sample set, not as a formal information-bottleneck objective or a proof of necessity.

3 Human-Centric Multi-Stage Evaluation

3.1 Three-Stage, Four-Dimension Framework

Figure 1 presents our human-centric process-safety framework for intersection driving. Given an intersection scenario x, it first decomposes the ego trajectory into three semantic stages: *approach*, *in-intersection*, and *exit*. This temporal decomposition localizes when an unsafe response occurs. In the recorded study, approach ends when the ego first enters the intersection polygon, in-intersection ends when it exits the polygon, and exit spans the remaining evaluation horizon.

Each stage is assessed along four complementary dimensions: Human-Factor-Based Safe Driving Procedures (H-SDP) completion, risk, feasibility, and efficiency. H-SDP identifies whether the system completed the expected human-relevant driving procedure; risk captures proximity and severity of potential conflicts; feasibility checks whether behavior remains behaviorally and rule-consistent; and efficiency measures progress without overriding safety. Table 1 instantiates this common framework for the three stages. Together, the two axes

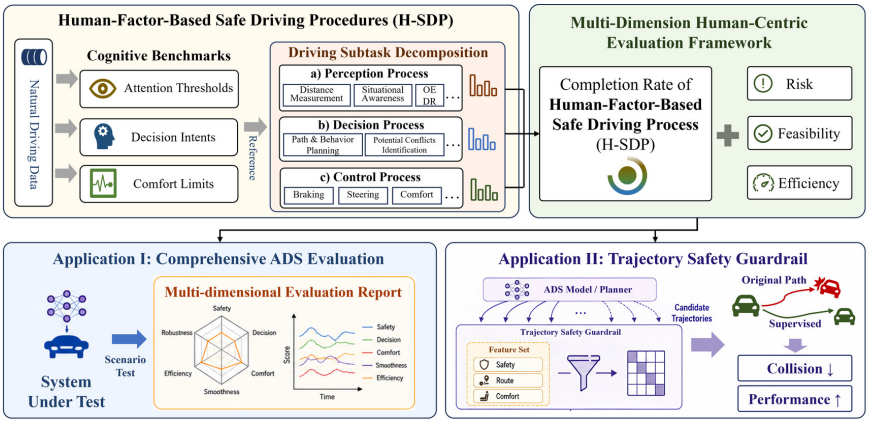


Fig. 1: Overview of the human-process-oriented framework. Human-factor-based safe driving procedures decompose intersection driving into perception, decision, and control obligations. The framework supports ADS diagnosis and trajectory safety evaluation.

Table 1: Stage-wise instantiation of the three-stage, four-dimension framework.

Stage	H-SDP completion	Risk	Feasibility / efficiency
Approach	lane, object, traffic-light, route, and deceleration subtasks	TTC	priority response, speed and headway control
In intersection	traffic-flow perception, conflict judgment, acceleration and steering control	MPrTTC, PCSI	right-of-way response, crossing speed, influence on others
Exit	turn-path monitoring, speed-limit judgment, comfort control	TTC	lane-line response, post-intersection speed and headway

distinguish a safe outcome from a safe driving process and support localized diagnosis rather than a single scenario-level label.

3.2 Human-Factor-Based Safe Driving Procedures

H-SDP operationalizes the human-relevant procedure expected at each stage of the framework. Figure 2 summarizes the decomposition into perception, decision, and control obligations. Let $T_i(\Omega, x, t) \in [0, 1]$ denote the completion score of subtask i by system Ω at time t. Our proposed evaluation protocol encompasses more than twenty operational checks. *Perception* covers lane recognition, object coverage in the human field of view, traffic-light recognition, and traffic-flow perception. *Decision* covers localization confidence, stage awareness, replanning, path feasibility, traffic-light response, risk awareness, and occlusion judgment.

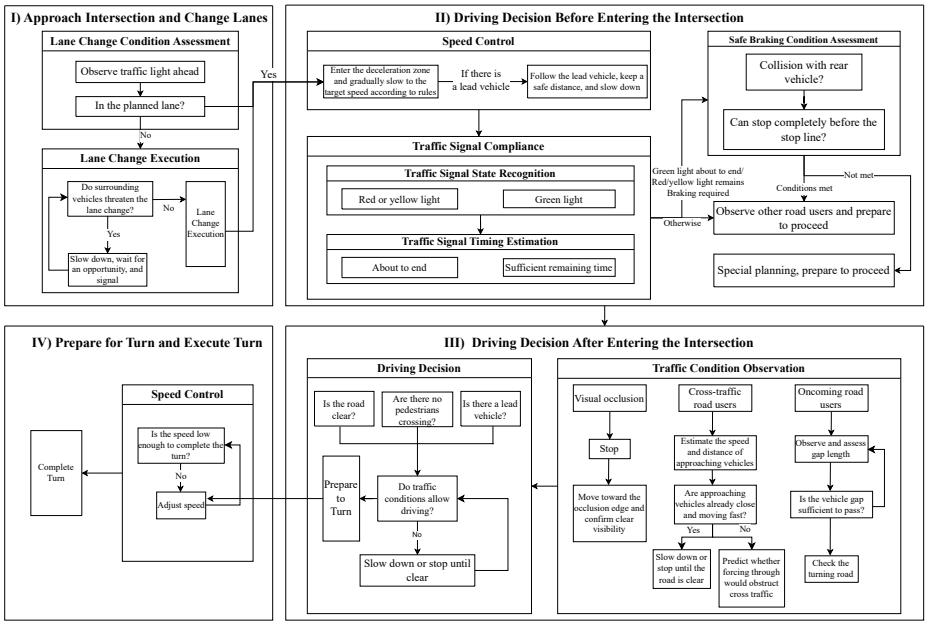


Fig. 2: Human driving-process decomposition. The stage split localizes when an unsafe response occurs, while the H-SDP graph specifies which perception, decision, or control subtask should be checked.

Control covers tracking frequency, acceleration, jerk, and lateral acceleration. Table 2 summarizes the primary check families comprising our taxonomy.

Subtasks include binary checks, such as whether the current lane is correctly detected, and ratio scores, such as whether detected objects cover the human driver's visible target set:

$$T_{\text{obj}}(\widehat{S}, S, R_p) = \frac{|S \cap \widehat{S} \cap R_p|}{|S \cap R_p| + \epsilon}, \tag{1}$$

where S is the human-visible target set, $\widehat{S}$ is the ADS output, R_p is the human field of view, and ϵ prevents division by zero. The stage-level human SDP completion is

$$C_s(\Omega, x) = \frac{1}{|\mathcal{T}_s| \sum_{i \in I_s} w_i} \sum_{t \in \mathcal{T}_s} \sum_{i \in I_s} w_i T_i(\Omega, x, t), \tag{2}$$

with task set I_s, time indices $\mathcal{T}_s$, and non-negative task weights w_i with $\sum_i w_i > 0$ (unit weights recover the unweighted average). This normalization ensures $C_s \in [0, 1]$.

Occlusion is treated as a decision obligation rather than a pure perception failure. Let $\widehat{B}$ be the set of planned path points, R_p the visible region, and $R_{pb} \subseteq$

Table 2: H-SDP check families in the proposed protocol. Binary, ratio, and thresholded comfort scores share a $[0, 1]$ interface; thresholds are configured per scenario based on typical human-driving comfort bounds.

Category	Subtask	Score type	Human-process interpretation
Perception	lane / map feature recognition	binary match	vehicle knows where it is allowed to drive
Perception	surrounding object coverage	visible-set recall	ADS should see at least what a human driver can see
Perception	traffic-light color	visible-set color recall	signal state is correctly perceived before entering
Decision	scenario-stage awareness	non-empty stage output	planner knows whether it is approaching, crossing, or exiting
Decision	replanning frequency	thresholded frequency	planner updates no slower than human tactical reaction
Decision	occlusion judgment	blocked-path slow-down check	high occlusion should trigger cautious low-speed behavior
Control	braking / acceleration comfort	threshold check	avoid harsh control unless safety-critical
Control	jerk / lateral acceleration	threshold check	passenger-comfort and controllability envelope

R_p its occluded subregion. If more than half of the visible planned points lie in R_{pb} and the ego speed remains above a cautious threshold v_{occ}, the occlusion-response score is zero:

$$T_{\text{occ}} = \mathbb{I}\left[\frac{|\widehat{B} \cap R_{pb}|}{|\widehat{B} \cap R_p| + \epsilon} \leq 0.5 \ \vee \ v \leq v_{\text{occ}}\right]. \tag{3}$$

This explicitly captures the defensive-driving rule that uncertainty should reduce speed before the vehicle commits to the conflict area.

3.3 Risk, Feasibility, and Efficiency Metrics

Risk is reported with TTC for near-collision checks and MPrTTC for predicted intersection conflicts. We additionally use the potential collision severity index (PCSI), which accounts for conflict direction and relative speed [33]. Feasibility records right-of-way, stop-line, crosswalk, occlusion-caution, and smooth-control compliance. Efficiency is measured from delay and forward progress and is explicitly subordinate to process safety.

To streamline compact scenario-level reporting, H-SDP completion, feasibility and efficiency metrics are discretized into five ordinal rating bands. For risk assessment, a scenario is deemed acceptable only if its maximum observed risk level does not exceed the predefined human reference agent for that scenario.

4 Experiment I: Full-Stack Evaluation on Autoware

4.1 Setup

We evaluate Autoware.ai 0.14.1 [16] in CARLA [5] Town05. Autoware receives LiDAR, RGB, IMU, map, and localization inputs through ROS bridge topics and returns longitudinal/lateral control commands. The scenario set contains 10 single-conflict cases (D1–D10) and 5 multi-conflict cases (M1–M5). The multi-conflict cases combine events such as motorcycle cut-ins, pedestrian crossings, cyclist crossings, right-of-way conflicts, and adverse visibility, making them useful for testing whether a system can preserve process safety across repeated local disturbances.

(a) D10 rainy straight-through conflict.

(b) M1 multi-conflict left turn.

Fig. 3: Representative scenario snapshots extracted from the CARLA–Autoware SIL runs. D10 is an outcome success with a near-miss process; M1 is a dense failure case with multiple conflict sources.

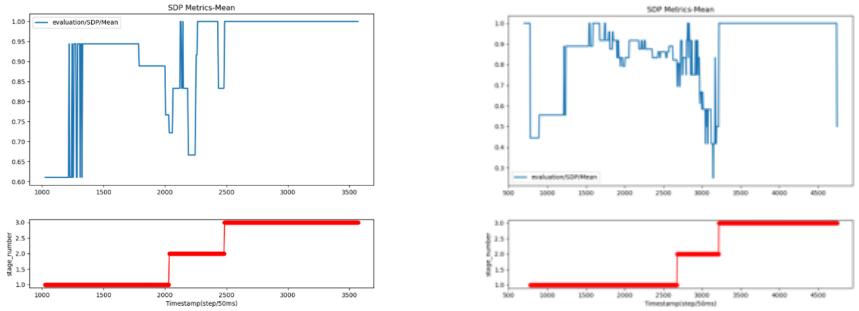
4.2 Case Studies

D10: success but unsafe. In D10, the ego vehicle drives straight through a rainy intersection while an oncoming vehicle incorrectly takes priority and turns left across the ego path (Fig. 3). Autoware avoids collision through a small steering adjustment, so an outcome-only test would mark the case as passed. Our process metrics give a different diagnosis: during the in-intersection stage, the decision subtask drops because the ego vehicle does not proactively decelerate at the conflict point. The minimum MPrTTC reaches 0.01 s. The global SDP completion is 88.39%, but the stage-wise trace reveals that this average hides a short high-risk interval. This is the central SOTIF failure mode targeted by the framework: no collision occurs, but the process is not acceptably safe.

M1: dense multi-conflict failure. In M1, the ego vehicle turns left in clear weather, encounters a motorcycle cut-in/cut-out, passes a pedestrian near the stop line without yielding, collides with a crossing cyclist when exiting, and then brakes for a later motorcycle conflict. The SDP trace falls before the cyclist collision and reaches a minimum of 0.2. PCSI and MPrTTC peak at the collision moment and also capture secondary pedestrian-crossing risk. The feasibility trace drops to zero when the ego vehicle fails to yield at the vulnerable-road-user conflict, while traffic-impact metrics show denser low-TTC interactions than D10. The failure is therefore not only a collision label; it is a sequence of perception-side vulnerability and control/priority errors.

4.3 Scenario-Library Results

Across 15 intersection scenarios, Autoware passes 53.3% of cases. Its mean H-SDP completion is 83.22%, the mean reported minimum risk time is 0.20 s, and mean feasibility is 77.45%. The finite-sample mean speed and delay are 7.36 m/s and -0.135 s, respectively. Failed cases are dominated by perception weaknesses, especially lateral perception in pedestrian crossing and emergency cut-in situations. Passed cases still show control weaknesses, with high acceleration or hard



(a) D10 SDP trace. **(b)** M1 SDP trace.

Fig. 4: Process metrics expose risk that is not captured by pass/fail alone. D10 is collision-free but the in-intersection SDP score drops when the ego vehicle does not slow for an oncoming left-turn conflict. M1 fails after repeated conflicts and reaches an SDP minimum near 0.2 at the cyclist collision.

braking. The system also tends to prioritize rapid passage over right-of-way compliance, which explains why efficiency is high while process safety remains insufficient.

Figure 5 visualizes the corresponding mean ordinal ratings. It highlights high intersection efficiency, weak risk benchmarking, and intermediate H-SDP completion and intersection compliance; the rating definitions are given in Appendix A.

Table 3 shows that low reported risk times are not confined to failed cases. D8–M4 and D10 all record a 0.01 s minimum risk time, marking critical interaction events. Conversely, feasibility and efficiency do not move together: M1 is fast (negative delay) but unsafe, while D10 has positive delay yet still has a severe near miss. These patterns support using a multi-dimensional report rather than a single scalar score. They also identify improvement targets: lateral perception for vulnerable road users, smoother control in passed cases, and right-of-way logic when the ego vehicle has a mission to clear the intersection quickly.

5 Experiment II: Offline Candidate-Selection Case Study

5.1 Setup

We test a process-inspired selector on a fixed pool of OpenDriveVLA [36] candidates from a near-balanced nuScenes intersection subset: five groups of 100 samples and one red-right-turn/stop group of 4 samples. Each of the 504 samples has 8 sampled future ego trajectories, yielding 4032 candidates. The first serialized candidate, the full selector, an oracle risk selector using future information, and compact tested variants are evaluated on this same pool.

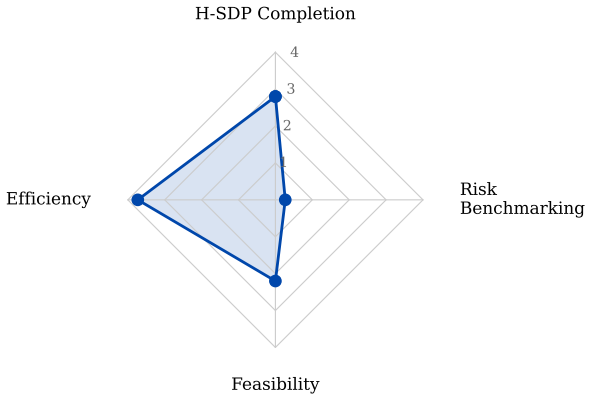


Fig. 5: Radar-chart visualization of the scenario-library mean ordinal ratings on a 0–4 scale. Intersection compliance denotes the feasibility dimension; the rating definitions and aggregation are given in Appendix A.

Fig. 6: Offline candidate-selection example. The selector re-ranks fixed-pool trajectories using process-inspired costs.

For each fixed candidate set $\{\tau_k\}_{k=1}^{K}$, we select the trajectory with the lowest weighted process-inspired cost:

$$k^* = \arg\min_k \lambda_{\text{safe}} J_{\text{safe}}(\tau_k) + \lambda_{\text{route}} J_{\text{route}}(\tau_k) + \lambda_{\text{comfort}} J_{\text{comfort}}(\tau_k) + \lambda_{\text{eff}} J_{\text{eff}}(\tau_k) + \lambda_{\text{rule}} J_{\text{rule}}(\tau_k). \tag{4}$$

The costs cover safety, route consistency, comfort, efficiency, and rule/map compliance. The selector does not modify OpenDriveVLA weights or train a new planner. It uses current/history ego motion, route command, intersection stage, map metadata, and current agent boxes and velocities; Table 4 lists the signal groups.

5.2 Results

The full process-inspired selector reduces collision steps from 19 to 12 and UniAD/STP-3 collision averages from 0.79/0.23 to 0.53/0.13. These results in-

Table 3: Full scenario-library results from the SIL evaluation. "Key fail" is the dominant failed H-SDP subtask family. "Min risk time" is the minimum TTC or MPrTTC value. The final row reports the pass rate over all scenarios and means over finite observations for the remaining numeric columns; NAN/INF mark cases in which the ego does not complete the intersection.

ID	Pass	SDP (%)	Key fail	Min risk time (s)	Feas. (%)	Mean speed (m/s)	Delay (s)
D1	yes	89.13	Control	0.69	100.00	7.8925	0.127
D2	yes	87.22	Control	0.53	100.00	8.3548	-0.016
D3	no	78.39	Perception	0.03	100.00	7.9371	0.112
D4	yes	84.33	Control	0.27	77.23	7.6932	0.194
D5	yes	89.62	Control	0.71	75.12	7.7053	0.190
D6	yes	90.02	Control	0.43	100.00	7.6392	0.213
D7	yes	85.28	Control	0.24	73.82	7.2357	0.362
D8	no	81.21	Perception	0.01	72.65	7.0275	0.445
D9	no	77.52	Perception	0.01	NAN	–	INF
D10	yes	88.39	Control	0.01	82.32	6.5345	0.267
M1	no	81.93	Control	0.01	86.67	7.7614	-1.192
M2	no	73.88	Control	0.01	30.21	6.4219	-0.330
M3	no	78.90	Control	0.01	69.32	5.2925	-2.237
M4	no	80.18	Perception	0.01	39.68	8.3870	0.058
M5	yes	82.30	Control	0.08	77.26	7.2124	-0.081
Mean	53.3%	83.22	–	0.20	77.45	7.36	-0.135

Table 4: Signal groups used by the offline candidate selector.

Group	Features
Safety	clearance, TTC, MPrTTC gap, earliest overlap, surrounding-safety cost
Interaction	front/rear TTC, turn-conflict gap, vulnerable-path clearance
Route	route cost, final lateral/forward displacement, final heading
Efficiency	forward progress, mean speed, expected-progress cost
Comfort	max speed, acceleration, deceleration, jerk, lateral acceleration
Rule/Map	off-road, stop/turn, lane-keeping, intersection-caution, traffic-light, stop-line costs

dicate that while end-to-end models can generate diverse trajectories, their default imitation-based selection often misses the most defensive option. Within this fixed candidate pool, process-inspired costs select candidates with lower collision metrics without requiring upstream retraining.

The compact three-group variant agrees with the full selector on 71.43% of samples. It retains 12 collision steps. Its STP-3 collision average is 0.15, while its UniAD collision average (0.66) is worse than the full selector (0.53). Adding Rule/Map raises full-selector agreement to 82.34% but worsens collision steps to 14 and STP-3 collision to 0.17.

Interestingly, enforcing rigid map and rule compliance slightly increases collision occurrences in this test set. This phenomenon reveals a practical conflict between absolute safety and traffic-rule adherence: in emergency interactions,

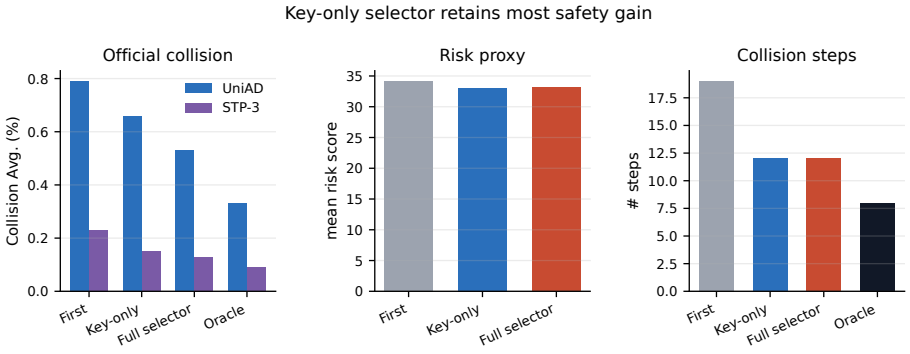


Fig. 7: Key-only validation. The tested Safety+Route+Comfort variant preserves the collision-step result but does not match the full selector's UniAD collision score.

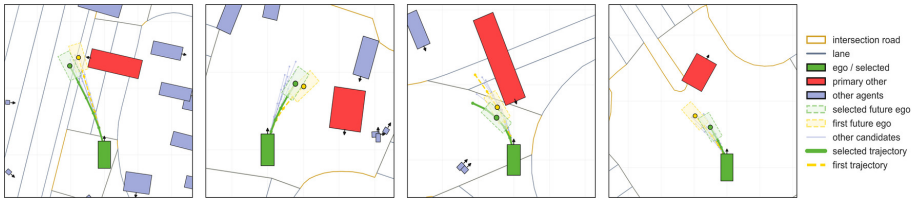


Fig. 8: Four representative fixed-pool cases from Experiment II. Each panel contrasts the first candidate (yellow dashed trajectory) with the process-inspired selected candidate (green trajectory) in the presence of the primary interacting agent (red). Pale trajectories denote the remaining candidates.

strict compliance, such as maintaining the lane center, may constrain necessary evasive maneuvers. A robust safety guardrail must therefore dynamically relax nominal rule constraints when imminent physical risk is detected.

5.3 Exploratory Feature-Group Sensitivity

We apply four post-hoc probes on the same 504 samples: Fisher-like sensitivity for selector decisions, permutation importance for selector decisions, permutation importance for safety improvement, and leave-one-group-out change. Table 6 reports their normalized scores.

The compact three-group variant agrees with the full selector on 71.43% of samples. It retains 12 collision steps. Its STP-3 collision average is 0.15, while its UniAD collision average (0.66) is worse than the full selector (0.53). Adding Rule/Map raises full-selector agreement to 82.34% but worsens collision steps to 14 and STP-3 collision to 0.17.

Table 5: Fixed-pool, single-run trajectory-selection results on 504 intersection samples. UniAD/STP-3 collision rates are evaluator averages (lower is better); "Risk" is the mean proxy score; "Coll. steps" is the count of collision-marked evaluation steps.

Method	UniAD Coll.	STP-3 Coll.	Risk	Coll. steps
First candidate	0.79	0.23	34.148	19
Full selector	0.53	0.13	33.145	12
Oracle risk selector	0.33	0.09	32.794	8
Key-only: Safety+Route+Comfort	0.66	0.15	33.067	12
+ Rule/Map	0.73	0.17	32.943	14

Table 6: Exploratory, in-sample feature-group sensitivity scores. Columns are complementary probes rather than a single ranking objective; high agreement with selector decisions does not necessarily imply better stored collision metrics.

Group	Fisher-like selector	Selector permutation	Safety-improvement permutation	Leave-one-group-out
Safety	0.0078	0.1685	0.1523	0.0737
Interaction	0.0013	0.0411	0.1557	0.1407
Route	0.1969	0.2337	0.3329	0.2211
Efficiency	0.6879	0.0372	0.0409	0.1524
Comfort	0.0937	0.4879	0.2899	0.1742
Rule/Map	0.0123	0.0316	0.0283	0.2379

6 Discussion and Conclusion

This paper argues that relying solely on outcome-oriented metrics, such as collision rates or minimum TTC, is insufficient for a comprehensive Safety of the Intended Functionality (SOTIF) evaluation. Such metrics often obscure whether an autonomous driving system is driving defensively or merely escaping collisions through last-second emergency actions. To address this, we introduced a human-centric, multi-stage evaluation framework that decomposes intersection scenarios into approach, in-intersection, and exit stages. By assessing each stage across four dimensions—safe driving procedure (H-SDP) completion, risk, feasibility, and efficiency—our framework provides a high-resolution diagnosis of system behavior.

Our closed-loop experiments with Autoware demonstrated the framework's ability to expose hidden process risks in otherwise "successful" collision-free runs and to pinpoint specific algorithmic vulnerabilities, such as lateral perception failures and right-of-way logic deficits. Furthermore, the offline trajectory selection study revealed that process-oriented metrics can be repurposed as a trajectory guardrail. By simply applying human-process-inspired costs, we successfully re-ranked end-to-end planner candidates and reduced collision occurrences without retraining the underlying model.

Limitations and Future Work. First, the current trajectory selector operates offline on a fixed candidate pool; transitioning this into a computationally lightweight, online safety filter requires further optimization of the metric calculation pipeline. Second, the weights of the process-inspired costs are currently

static. As observed in our feature sensitivity analysis, rigid adherence to rules can sometimes conflict with emergency collision avoidance. Future work will explore dynamic weight adaptation mechanisms that shift priority from rule compliance and comfort to absolute physical safety as interaction criticality increases. Finally, expanding the human-centric subtasks to account for diverse regional driving cultures will further enhance the framework's generalizability.

A Radar-Chart Rating Definitions

Figure 5 visualizes mean *ordinal ratings*, rather than the raw continuous values in Table 3. The rating rules are standardized as follows. Let $c(\Omega, x) \in [0, 1]$ be the scenario-level H-SDP completion of system Ω in scenario x, and let $l(\Omega, x) \in [0, 1]$ be its feasibility (intersection-compliance) score. Both use the same five-level map:

$$L(q) = \begin{cases} 4, & 0.90 < q \le 1, \\ 3, & 0.80 < q \le 0.90, \\ 2, & 0.70 < q \le 0.80, \\ 1, & 0.60 < q \le 0.70, \\ 0, & q \le 0.60, \end{cases} \qquad L_{\text{H-SDP}} = L(c), \quad L_{\text{feas}} = L(l). \tag{5}$$

Risk benchmarking is binary. Let $r^*(\Omega, x)$ denote the worst-case risk metric and Ω_h the human reference agent. Using the comparator's safety-oriented direction, the scenario receives a pass only when its worst case is no worse than the reference:

$$L_{\text{risk}}(\Omega, x) = \begin{cases} 1, & r^*(\Omega, x) > r^*(\Omega_h, x), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Thus, the risk axis is the mean binary pass score, not a calibrated continuous collision-risk estimate.

For intersection efficiency, let T_{actual} and T_{ideal} be the actual and ideal in-intersection traversal times. The relative delay is

$$m(\Omega, x) = \frac{T_{\text{actual}} - T_{\text{ideal}}}{T_{\text{ideal}}}, \qquad L_{\text{eff}}(\Omega, x) = \begin{cases} 4, & m \le 0.10, \\ 3, & 0.10 < m \le 0.30, \\ 2, & 0.30 < m \le 0.60, \\ 1, & 0.60 < m \le 1.00, \\ 0, & m > 1.00. \end{cases} \tag{7}$$

Negative delay is therefore assigned level 4 because it satisfies $m \le 0.10$ under this definition.

For each dimension d, the radar chart plots the mean scenario rating

$$\overline{L}_d = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} L_d(\Omega, x). \tag{8}$$

For the 15-scenario library, the displayed values are $\overline{L}_{\text{H-SDP}} = 2.8$, $\overline{L}_{\text{risk}} = 0.267$, $\overline{L}_{\text{feas}} = 2.2$, and $\overline{L}_{\text{eff}} = 3.73$. All axes use the common 0–4 display scale; because L_{risk} is binary, its mean lies in $[0, 1]$.

References

1. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
3. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A.H., Fletcher, L., Beijbom, O., Omari, S.: nuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–11 (2021)
4. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8748–8757 (2019)
5. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Proceedings of the Conference on Robot Learning. pp. 1–16 (2017)
6. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors **37**(1), 32–64 (1995)
7. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012)
9. Gipps, P.G.: A behavioural car-following model for computer simulation. Transportation Research Part B **15**(2), 105–111 (1981)
10. Hayward, J.C.: Near-miss determination through use of a scale of danger. In: Highway Research Record. vol. 384, pp. 24–34 (1972)
11. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: Proceedings of the European Conference on Computer Vision. pp. 533–549 (2022)
12. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., Li, H.: UniAD: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17853–17862 (2023)
13. International Organization for Standardization: ISO 26262:2018 Road vehicles – Functional safety (2018), international Standard
14. International Organization for Standardization: ISO 21448:2022 Road vehicles – Safety of the intended functionality (2022), international Standard
15. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: VAD: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)

16. Kato, S., Tokunaga, S., Maruyama, Y., Maeda, S., Hirabayashi, M., Kitsukawa, Y., Monrroy, A., Ando, T., Fujii, Y., Azumi, T.: Autoware on board: Enabling autonomous vehicles with embedded systems. In: Proceedings of the ACM/IEEE International Conference on Cyber-Physical Systems. pp. 287–296 (2018)
17. Kesting, A., Treiber, M., Helbing, D.: General lane-changing model MOBIL for car-following models. Transportation Research Record **1999**(1), 86–94 (2007)
18. Koopman, P., Wagner, M.: Challenges in autonomous vehicle testing and validation. In: SAE World Congress (2016)
19. Krajzewicz, D., Erdmann, J., Behrisch, M., Bieker, L.: Recent development and applications of SUMO – simulation of urban mobility. International Journal on Advances in Systems and Measurements **5**(3–4), 128–138 (2012)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774 (2017)
21. Mahmud, S.M.S., Ferreira, L., Hoque, M.S., Tavassoli, A.: Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. IATSS Research **41**(4), 153–163 (2017)
22. Menzel, T., Bagschik, G., Maurer, M.: Scenarios for development, test and validation of automated vehicles. In: IEEE Intelligent Vehicles Symposium. pp. 1821–1827 (2018)
23. Michon, J.A.: A critical view of driver behavior models: What do we know, what should we do? In: Human Behavior and Traffic Safety, pp. 485–524. Springer (1985)
24. Nistér, D., Lee, H.T.D., Ng, J., Wang, Y.: Safety force field. arXiv preprint arXiv:1911.06663 (2019)
25. Prakash, A., Chitta, K., Geiger, A.: Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. In: Conference on Robot Learning. pp. 182–201 (2021)
26. Rong, G., Shin, B.H., Tabatabaee, H., Lu, Q., Lemke, S., Možeiko, M., Boise, E., Uhm, G., Gerow, M., Mehta, S., Agafonov, E., Kim, T., Sterner, E.: LGSVL simulator: A high fidelity simulator for autonomous driving. In: IEEE International Conference on Intelligent Transportation Systems. pp. 1–6 (2020)
27. Shalev-Shwartz, S., Shammah, S., Shashua, A.: On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374 (2017)
28. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Luo, P., Geiger, A., Li, H.: DriveLM: Driving with graph visual question answering. arXiv preprint arXiv:2312.14150 (2023)
29. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2446–2454 (2020)
30. Treiber, M., Hennecke, A., Helbing, D.: Congested traffic states in empirical observations and microscopic simulations. Physical Review E **62**(2), 1805–1824 (2000)
31. Ulbrich, S., Menzel, T., Reschka, A., Schuldt, F., Maurer, M.: Defining and substantiating the terms scene, situation, and scenario for automated driving. In: IEEE International Conference on Intelligent Transportation Systems. pp. 982–988 (2015)
32. Vogel, K.: A comparison of headway and time to collision as safety indicators. Accident Analysis and Prevention **35**(3), 427–433 (2003)

33. Wang, H., Huang, Y., Khajepour, A., Zhang, Y., Rasekhipour, Y., Cao, D.: Crash mitigation in motion planning for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems **20**(9), 3313–3323 (2019). https://doi.org/10.1109/TITS.2019.2893622
34. Westhofen, L., Neurohr, C., Koopmann, T., Butz, M., Schütt, B., Utesch, F., Kramer, B., Gutenkunst, C., Böde, E.: Criticality metrics for automated driving: A review and suitability analysis of the state of the art. Archives of Computational Methods in Engineering **30**, 1–35 (2023)
35. Wu, P., Jia, X., Chen, L., Yan, J., Li, H., Qiao, Y.: TCP: Trajectory-guided control prediction for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 824–833 (2022)
36. Zhou, X., Han, X., Yang, F., Ma, Y., Knoll, A.C.: OpenDriveVLA: Towards end-to-end autonomous driving with large vision-language-action model. arXiv preprint arXiv:2503.23463 (2025), accepted to AAAI 2026