HIDRIVE: A CLOSED-LOOP BENCHMARK FOR HIGH-LEVEL AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end autonomous driving has witnessed rapid progress, yet existing benchmarks are increasingly saturated, with state-of-the-art models achieving near-perfect scores on widely used open-loop and closed-loop benchmarks. This saturation does not mean that the problem has been solved; instead, it reveals that current benchmarks remain limited in scenario diversity, object variety, and the breadth of driving capabilities they evaluate. In particular, they lack sufficient long-tail scenarios involving rare but safety-critical objects and fail to assess advanced decision-making such as legal compliance, ethical reasoning, and emergency response. To address these gaps, we propose HiDrive, a new closed-loop benchmark for end-to-end autonomous driving that emphasizes long-tail scenarios and richer capability evaluation. HiDrive introduces a diverse set of rare objects and uncommon traffic situations, enabling evaluation not only of fundamental driving skills but also of performance in long-tail scenarios through a richer set of capability assessments, including rule compliance, moral reasoning, and context-dependent emergency maneuvers. Correspondingly, we extend previous collision-avoidance-centered metrics into a comprehensive evaluation system that encompasses collision and braking, traffic-rule compliance, and moral-reasoning indicators. Built on a more advanced physics engine, HiDrive provides physically realistic lighting and high-fidelity visual rendering, offering a more challenging and realistic testbed for assessing whether autonomous driving systems can handle the complexity of real-world deployment. The HiDrive software, source code, digital assets, and documentation are available at https://huggingface.co/HiDriveBenchmark/HiDriveBenchmark.

1 INTRODUCTION

Autonomous driving, which aims to transform raw sensor observations into safe and reliable driving decisions, has made remarkable progress in recent years and attracted increasing attention from both academia and industry. To measure such progress, autonomous driving benchmarks have gradually evolved from open-loop evaluation, which typically relies on pre-recorded real-world driving data and compares model outputs with human demonstrations, to closed-loop evaluation, which is usually conducted in simulation environments where the ego vehicle's actions can influence subsequent states and interactions. Compared with open-loop evaluation, closed-loop evaluation better reflects real-world deployment, because autonomous driving is inherently interactive and sequential. Additionally, in a closed-loop simulator, it is possible to construct various long-tail scenarios that are difficult to collect from real-world data.

However, the rapid improvement of recent models has led to clear saturation on these benchmarks, where state-of-the-art methods achieve increasingly high scores, leaving little room for meaningful differentiation. For example, on the most widely used open-loop benchmarks, including nuScenes Caesar et al. (2020), NAVSIM-v1 Dauner et al. (2024), and NAVSIM-v2 Dauner et al. (2024); Cao et al. (2025), state-of-the-art methods have achieved a collision rate of 0.03% Xia et al. (2025a), 94.8% Best-of-N PDMS Chen et al. (2026), and 89.3% EPDMS Wang et al. (2026), respectively. On Bench2Drive Jia et al. (2024), one of the most commonly used closed-loop benchmarks, state-of-the-art methods have reached a driving score of 89.2% Sun et al. (2026). This saturation does not necessarily indicate that end-to-end autonomous driving has been solved; rather, it reveals that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

existing benchmarks remain relatively simple in terms of scenario complexity, object diversity, and the range of driving abilities evaluated.

First, current benchmarks include only a limited number of long-tail scenarios, despite their being crucial for safe real-world driving. In daily traffic environments, autonomous vehicles may encounter rare but safety-critical objects, such as various accident warning signs, traffic cones, damaged or abandoned public facilities, or irregularly shaped obstacles. These objects may appear infrequently in standard datasets, but failing to correctly perceive or reason about them can lead to severe consequences. Existing benchmarks are often dominated by common categories such as cars, pedestrians, cyclists, and traffic signs, which encourages models to optimize for frequent cases while leaving their robustness to rare objects largely underexplored. As a result, high benchmark performance may mask serious weaknesses in handling long-tail risks.

Second, existing benchmarks mainly focus on basic driving abilities, such as lane following, obstacle avoidance, traffic light recognition, and collision-free navigation. However, real-world driving requires much richer and more nuanced decision-making. For example, when a vehicle passes through a waterlogged road section with pedestrians nearby, it should slow down to avoid splashing water onto them. Similarly, when an ambulance or police car approaches from behind while the ego vehicle is waiting at a red light, the vehicle may need to make an emergency maneuver to yield, even if this requires temporarily entering an otherwise prohibited area or crossing the stop line under exceptional circumstances. These scenarios go beyond simple rule-following: they require an understanding of traffic laws, social norms, moral principles, and the hierarchy of safety-critical decisions. Therefore, a comprehensive autonomous driving benchmark should evaluate not only whether a vehicle can drive safely in common cases, but also whether it can reason appropriately in complex, ambiguous, and norm-sensitive situations.

To address these challenges, we propose HiDrive, a new closed-loop benchmark for end-to-end autonomous driving that emphasizes long-tail scenarios and a richer evaluation of driving capabilities. Our benchmark introduces a diverse set of rare objects and uncommon traffic situations that are underrepresented in existing planning benchmarks. In addition, it expands the scope of evaluation from basic driving skills to more advanced capabilities, including legal compliance, ethical reasoning, emergency response, and context-dependent decision-making. Accordingly, we extend the previous collision-avoidance-centered evaluation metrics to a comprehensive evaluation system that covers collision and braking, traffic rule compliance, and moral reasoning indicators. Additionally, we adopt a more advanced physics engine to support physically plausible dynamics, realistic lighting, and high-fidelity visual rendering. By doing so, the proposed benchmark provides a more challenging and realistic testbed for assessing whether autonomous driving systems can handle the complexity of real-world deployment.

2 Related Work

2.1 Planning Benchmarks for Autonomous Driving

Autonomous driving benchmark evaluation is gradually evolving from open-loop datasets to integrated planning benchmarks. In recent years, research has increasingly focused on evaluating the quality of planning for learning-based end-to-end systems under interactive protocols. Table 1 compares representative benchmarks from the perspectives of closed-loop property, ability/scenario scale, testing-route volume, and legal/ethical coverage.

Table 1: **Planning benchmark comparison.** We compare benchmarks across closed-loop property, ability/scenario scale, test-route quantity, and legal/ethical coverage.

Benchmark	Closed-Loop	Abilities	Scenarios	Test Routes	Legal Coverage	Ethical Coverage
nuScenes	✗	–	1,000	–	Collision-only	✗
NAVSIM (v1/v2)	✗	–	–	–	Collision-only	✗
CARLA Leaderboard v2.0	✓	6	21	–	Collision, Red-light	Yielding
Bench2Drive	✓	5	44	220	Collision, Red-light, Speed-limit	Yielding
HiDrive (Ours)	✓	**30**	**94**	**330**	**More comprehensive**	**Comprehensive**

Existing benchmarks still have clear limitations, mainly due to their task definitions and evaluation boundaries. nuScenes Caesar et al. (2020) is a typical open-loop benchmark. Its strength lies in large-scale annotation and perception/prediction alignment. However, its evaluation is based on offline logs, and policy actions do not feed back to change other traffic participants. As a result, it cannot cover the full closed-loop process of "decision-environment feedback-redecision." For sudden, dynamic situations (e.g., a pedestrian appearing from behind an occlusion), open-loop metrics can serve as a reference but cannot directly replace closed-loop conclusions.

NAVSIM Dauner et al. (2024); Cao et al. (2025) extends log-driven simulation to a larger-scale, unified evaluation layer, offering a practical balance between cost and coverage. However, its "non-reactive" setup means background traffic mainly follows predefined trajectories, so tested policies have limited influence on other agents' behaviors. This makes NAVSIM suitable for large-scale screening and iteration, but less suitable for evaluating strong interaction capabilities with other participants (e.g., aggressive cut-ins or continuous yielding).

CARLA Leaderboard v2.0 Dosovitskiy et al. (2017) and Bench2Drive Jia et al. (2024) both provide sensor-level closed-loop evaluation, allowing direct assessment of route completion under infraction penalties. Their main limitations are twofold. First, scores are mostly route-aggregated, ability decomposition remains relatively coarse, and evaluation signals are relatively single, mainly collision-centered safety penalties and traffic-violation-centered responsibility penalties. It is therefore difficult to measure whether a model both avoids over-interfering with surrounding traffic and identifies potential risks early. Second, the simulation world still differs significantly from the real world, including limited asset diversity, constrained rendering realism, and insufficient coverage of long-tail behaviors (e.g., rule-breaking or abnormal traffic participants). Meanwhile, as of May 2026, the official CARLA AD Leaderboard lists v2.1 as the current version, keeps older versions separately supported, and states that the 2025 CARLA AD Challenge did not run Dosovitskiy et al. (2017).

In contrast, HiDrive further strengthens end-to-end evaluation through richer long-tail scenarios, higher-level, more comprehensive ability design, and a finer-grained ability-decomposition metric system. HiDrive defines 30 high-level ability categories, each integrating multiple fine-grained scenarios (e.g., left and right turns) into a single driving ability (e.g., turning), enabling a more comprehensive and detailed assessment of model driving ability.

2.2 End-to-End Autonomous Driving

Autonomous driving systems encompass a variety of paradigms, including traditional modular systems, end-to-end models, vision-language-action (VLA) models, world-action models (WAMs), and others. The end-to-end family has yielded a long line of architectural innovations. ST-P3 Hu et al. (2022) pioneered the unification of the three classical stages within a single trainable network. UniAD Hu et al. (2023) extends this concept around a BEV feature volume, propagating information through task-specific decoders in a style similar to BEVFormer Li et al. (2022). VAD Jiang et al. (2023) replaces dense spatial maps with a fully vectorized scene representation, while PARA-Drive Weng et al. (2024) investigates how to structure multi-task supervision and inter-module dependencies inside end-to-end stacks. GenAD Zheng et al. (2024) retains a BEV backbone but distills sparse detection and mapping features for planning, whereas SparseDrive Sun et al. (2025) directly extracts those sparse features from the image stream to reduce error accumulation. DiffusionDrive Liao et al. (2025) samples trajectory candidates using a truncated diffusion process. HENet++ Xia et al. (2025b) jointly extracts sparse foreground instances and dense panoramic BEV features via hybrid attention. TCP Wu et al. (2022) introduces a dual-path design that fuses trajectory and control signals to boost planning accuracy. ThinkTwice Jia et al. (2023b) proposes a two-phase planner that first generates candidate trajectories and then refines them using scene context, implicitly capturing a world-model intuition. DriveAdapter Jia et al. (2023a) distills knowledge from an expert planner to improve planning robustness. DriveTransformer Jia et al. (2025) employs parallel queries for sparse sensor interaction in a streaming multi-task learning setup. Recent language- and world-model-based methods further couple multimodal reasoning with driving control. DriveLM (Sima et al., 2024) frames driving-relevant relations as graph-based visual question answering. Reason2Drive (Nie et al., 2024) enables chain-style interpretable reasoning, and DriveGPT4 (Xu et al., 2024) formulates end-to-end driving through a large language model. DriveAgent (Hou et al., 2025) combines multi-agent structured reasoning with multimodal sensor fusion, while CoMAL (Yao et al., 2025) studies collaborative reasoning across mixed-autonomy traffic. ORION (Fu et al., 2025) pairs

a VLM with end-to-end driving for command generation, and SimLingo (Renz et al., 2025) aligns natural-language maneuver prompts with action sequences for language-conditioned control. Know-Val (Xia et al., 2025a) combines retrieval-guided open-world perception with knowledge retrieval and selects trajectories through a learned value model over world-model predictions. Curious-VLA (Chen et al., 2026) improves exploration in driving VLA policies, and Latent-WAM (Wang et al., 2026) learns compact latent world-action representations for end-to-end planning.

3 HiDrive

3.1 Benchmark Overview

HiDrive is a closed-loop benchmark built on the CARLA simulator Dosovitskiy et al. (2017) using the CARLA UE5 branch for end-to-end autonomous driving evaluation. Compared with prior CARLA-based benchmarks such as Bench2Drive Jia et al. (2024), HiDrive adopts a newer rendering branch with more realistic lighting, texture, and scene details, which helps reduce the simulation-to-reality gap.



Figure 1: **Scenario comparison between HiDrive (Ours) and Bench2Drive.** Built with an updated physics-rendering engine and richer high-fidelity assets, HiDrive provides more realistic visual quality and lighting effects than existing simulation benchmarks such as Bench2Drive Jia et al. (2024).

The benchmark is organized as 330 short routes (about 150 meters per route) and covers 30 high-level scenario categories with 94 concrete scenario instantiations. Routes are distributed across diverse road topologies and traffic contexts, and each route contains a specified start point, end point, route layout, traffic participants, and potential obstacles. This design supports fine-grained capability diagnosis under controlled yet diverse closed-loop conditions.

3.2 Long-Tail Scenario Design

HiDrive follows a closed-loop, route-based task protocol and provides mainstream multimodal input interfaces consistent with prior CARLA-based benchmarks Jia et al. (2024); Dosovitskiy et al. (2017) (e.g., surround-view cameras, LiDAR, and optional radar), enabling low-cost migration and fair comparisons with existing methods. In scenario construction, while retaining the foundational scenarios and abilities covered by previous baselines such as Bench2Drive Jia et al. (2024), we emphasize complex and long-tail conditions that are likely to appear in real-world deployment but remain underrepresented in existing benchmarks. Under realistic human driving conventions and traffic rules, we extend scenario design in three aspects:

- **Richer instance variations.** In many prior methods, in-lane obstacle settings are mostly limited to obvious road markers, with few types and highly homogeneous appearances. For scenarios requiring obstacle handling, we keep the core task logic unchanged while introducing more obstacle categories and appearance variations, testing whether the policy can still perform robust recognition and decision-making under unfamiliar conditions.
- **More complex traffic compositions.** We increase the behavioral diversity of traffic participants, meaning that other agents in the same scene do not necessarily comply strictly with standard traffic rules. Irregular behaviors (e.g., unexpected stops or rule violations) are included to evaluate the model's robustness to non-ideal interactions.

- **Stronger social interaction constraints.** Beyond explicitly avoiding collisions and obeying traffic signals, we introduce more socially grounded interaction scenarios. For example, when passing pedestrians near puddles, the ego vehicle should slow down appropriately to avoid splashing water. In these cases, the model must identify potential interaction objects in the scene and make reasonable ethical decisions.



A cyclist suddenly emerges from behind an obstruction.

A roadside vehicle suddenly opens its door; watch for collisions or passengers getting out.

A broken-down vehicle is stopped on the road; switch on the hazard lights and place a warning barrier.

Pedestrians are waiting near standing water; slow down and proceed carefully.

The vehicle is stopped by police and should pull over and comply.

Diverse road obstacles.

Figure 2: **Representative long-tail scenarios in HiDrive.** These examples cover sudden emergence, door-opening risk, puddle-side pedestrian interaction, police-stop compliance, and diverse uncommon obstacles, highlighting richer long-tail diversity for robustness and social-reasoning evaluation.

The design and selection of the above long-tail scenarios are guided by the KnowVal Xia et al. (2025a) knowledge graph, which covers traffic laws, defensive-driving principles, moral principles, and common-sense norms. These challenging events are embedded in ordinary routes that also test common driving skills. Because the event templates are modular, they can be relocated or recombined across compatible roads and intersections.

3.3 Ability-Oriented Protocol

Beyond route-level average scores, HiDrive evaluates model behavior through ability-oriented analysis. This ability-oriented view spans the driving spectrum in closed loop, from foundational driving skills to higher-order capabilities involving legal compliance, norm-aware interaction, and dilemma-level reasoning. According to scenario complexity and reasoning demand, we divide abilities into three sets:

- **Basic Set (basic operational ability).** This set includes 11 foundational and core autonomous-driving skills, such as yielding, detouring, and merging into traffic flow. At the ability level, it largely covers the core abilities in previous Bench2Drive Jia et al. (2024) and CARLA Leaderboard Dosovitskiy et al. (2017) settings, while extending long-tail object conditions; therefore, the overall difficulty is also increased.
- **Hard Set (rule- and norm-aware ability).** This set includes 10 abilities involving legal, ethical, and defensive-driving judgments, such as keeping a safer distance from abnormally driven vehicles. In these scenarios, the model must actively perceive the surrounding environment and reason to produce actions that ensure self-safety while conforming to ethical norms.
- **Thorny Set (dilemma-level decision ability).** This set contains 9 abilities that require handling rule conflicts or abnormal conditions, such as responding to emergency-vehicle pressure under red-light constraints or passing intersections with failed traffic lights. In some of these scenarios, violations or collisions may be unavoidable; the model needs stronger reasoning and decision-making to reduce loss, e.g., crossing a red light to clear the path

for an emergency vehicle, or moving off the road into open roadside space to avoid hitting pedestrians.

Table 2 presents the definitions and descriptions of the above 30 abilities and provides examples of corresponding scenarios. HiDrive builds an evaluation around these high-level ability categories, where each category contains multiple fine-grained scenarios (e.g., left turn and right turn). This design allows results to reflect overall performance while more clearly localizing model weaknesses at the ability level.

Table 2: **Ability taxonomy and concise descriptions.** We organize 30 abilities into Basic, Hard, and Thorny sets, enabling fine-grained diagnosis from routine maneuvers to legal-ethical judgment and dilemma-level decision making.

Ability	Description
Basic Set	
Emergency avoidance	Decelerate or change lanes when pedestrians/cyclists suddenly emerge from blind spots.
Obstacle detouring	Handling static obstacles such as roadwork zones or broken-down vehicles with hazards and warning barriers.
Signalized turning	Complete left/right turns at signalized intersections while coordinating with surrounding traffic flow.
Cut-in response	Respond safely to overtakes, lane changes, and roadside vehicles merging into the main lane.
Traffic merging	Start from curbside or merge from highways/roundabouts into moving traffic.
Constrained-segment passage	Pass narrow or partially blocked segments (e.g., cones occupying both sides of the lane).
Overtaking	Overtake very slow lead vehicles when road conditions and right-of-way permit.
U-turn execution	Perform compliant U-turns at legally permitted locations.
Narrow-road following	Maintain safe headway with front/rear vehicles in narrow road sections.
Reasonable speed keeping	Keep appropriate speed for road context, avoiding both underspeeding and overspeeding.
Oncoming encounter etiquette	During opposite-direction encounters, keep farther from the centerline and slow down appropriately.
Hard Set	
Pedestrian-related ethics	Respect nearby pedestrians, e.g., avoid high-speed passage through puddles that could splash them.
Special yielding scenarios	Yield beyond collision-only logic, such as giving way to emergency vehicles or anticipating door-opening risks near curbside parked cars.
Open-world detouring	Handle uncommon non-standard obstacles, e.g., collapsed public facilities or scattered debris.
Speed-bump handling	Cross speed bumps at appropriately low speed.
Yielding in tight conflicts	At narrow intersections/segments with opposing traffic, yield reasonably instead of forcing passage.
Defensive distancing from erratic drivers	Keep larger safety margins from vehicles with suspicious behavior (frequent lane-straddling, weaving, possible drunk driving).
Police-stop compliance	Pull over and stop when intercepted by police instead of continuing to drive away.
Adverse-weather handling	Drive cautiously at a lower speed under low-visibility weather, such as fog.
Ego-failure mitigation	When sudden ego faults occur (e.g., brake failure), choose actions that minimize total loss.
Defensive turning under occlusion	Reduce speed when turning through corners with severe visual blind zones.
Thorny Set	
Forced lane borrowing	Cross lane markings when the current lane is blocked (construction, broken truck, etc.) and borrowing is necessary.
Signal-failure intersection handling	At intersections with failed traffic lights, decide left/right turns based on surrounding traffic conditions.
Intrusive cut-in risk mitigation	For very close, high-speed cut-ins, perform emergency braking; if collision is unavoidable, minimize harm.
Accident-scene handling	React appropriately to crashes ahead (e.g., rear-end incidents) and re-plan passage.
Wrong-way vehicle avoidance	Anticipate high-speed wrong-way vehicles and evade via lane change or safe roadside space.
Red-light emergency yielding	Under red light with emergency-vehicle pressure from behind, decide whether to cross to clear the path safely.
Partial sensor-blindness handling	Continue safe driving when camera/radar regions are degraded (blackout, blur, heavy corruption).
Value-priority dilemma handling	In unavoidable-collision dilemmas, prioritize human safety (e.g., steer into barriers to avoid pedestrians).
Defensive distancing from unknown objects	When feasible, keep distance from uncertain road objects that may indicate risk (e.g., suspected fuel spill).

3.4 Evaluation Metrics

We report four core route-level metrics in this section: drive score DS, route completion RC, legal score LS, and ethics score ES.

Overall scoring algorithm. For each route i, we initialize $LS_i = 1$ and $ES_i = 1$, and denote route completion by $RC_i \in [0, 1]$. For RC_i, in most cases, we compute completion as the ratio between passed route waypoints and total route waypoints: $RC_i = N_{i,\text{pass}}/N_{i,\text{all}}$. For some special scenarios (e.g., Wrong-way vehicle avoidance), the primary objective is ego safety rather than completing the full route; once a required segment is completed, we set RC_i to 1 directly. Given triggered events $\mathcal{T}_i$, HiDrive updates factors as follows:
(i) fixed legal-traffic events update LS_i;
(ii) collision and lane-outside events also update LS_i;
(iii) ethics events update ES_i only;
(iv) min-speed events may further modulate LS_i when background-speed signals are available. The route-level composed score is

$$DS_i = RC_i \cdot LS_i \cdot ES_i. \tag{1}$$

Thus, all four route-level metrics DS_i, RC_i, LS_i, and ES_i are in $[0, 1]$. Appendix A.1 details the representative penalty coefficients, their selection principles, the brake-aware relief rule, and a sensitivity analysis with every coefficient fixed to 0.5.

Split-level reporting convention. For split s, let $\mathcal{I}_s$ be the set of routes with valid records and $N_s = |\mathcal{I}_s|$. We report

$$DS = \frac{1}{N_s}\sum_{i\in\mathcal{I}_s} DS_i, \quad RC = \frac{1}{N_s}\sum_{i\in\mathcal{I}_s} RC_i, \quad LS = \frac{1}{N_s}\sum_{i\in\mathcal{I}_s} LS_i, \quad ES = \frac{1}{|I_{E,s}|}\sum_{i\in I_{E,s}} ES_i. \tag{2}$$

Here, $I_{E,s} \subseteq \mathcal{I}_s$ is the subset of ethics-applicable routes. It denotes the subset of routes that contain ethics-evaluation events.

4 Experiments

4.1 Baselines

To evaluate the performance of public end-to-end autonomous driving baselines on the HiDrive benchmark, we include the following methods and follow the input conditions, training recipes, and inference configurations specified in their official papers and open-source implementations:

- **TCP** Wu et al. (2022): jointly predicts trajectory and control signals; input is front-view camera plus ego state with TP navigation conditioning.
- **UniAD-Base** Hu et al. (2023): an explicit perception-prediction-planning pipeline with Transformer queries; input is multi-view camera with NC navigation conditioning.
- **VAD** Jiang et al. (2023): a Transformer-query architecture with vectorized scene representation; input is multi-view camera with NC navigation conditioning.
- **ThinkTwice** Jia et al. (2023b): a hierarchical coarse-to-fine planner with expert-feature distillation; input is camera with TP navigation conditioning.
- **DriveAdapter** Jia et al. (2023a): decouples perception and planning and bridges them with adapters; input is camera plus LiDAR with TP navigation conditioning.
- **DriveTransformer** Jia et al. (2025): a unified multi-task Transformer with parallel queries; input is multi-view camera with NC navigation conditioning.
- **ORION** Fu et al. (2025): a vision-language E2E policy for driving instruction generation; input is camera-only modality with NC navigation-command conditioning.
- **DiffusionDrive** Liao et al. (2025): a truncated-diffusion planner for multimodal trajectory generation; the official implementation uses a ResNet backbone with real-time inference.

Table 3: **HiDrive full-eval results.** We report split-level DS, RC, LS, and ES for Overall and Basic/Hard/Thorny sets. All values are reported as percentages (%).

Method	Overall				Basic Set				Hard Set				Thorny Set			
	DS	RC	LS	ES	DS	RC	LS	ES	DS	RC	LS	ES	DS	RC	LS	ES
TCP Wu et al. (2022)	18.4	23.7	27.8	89.3	18.5	24.1	27.2	91.0	17.6	21.4	28.0	87.1	19.1	25.5	30.1	85.0
UniAD-Base Hu et al. (2023)	21.9	26.6	32.1	91.6	21.2	26.7	31.6	93.7	23.6	25.3	30.8	89.8	22.3	28.0	36.4	85.2
VAD Jiang et al. (2023)	20.1	24.3	30.7	89.3	19.2	24.2	30.0	91.2	22.4	24.6	31.1	87.3	20.0	24.6	33.0	84.2
ThinkTwice Jia et al. (2023b)	28.3	35.6	38.9	92.4	26.3	36.8	38.3	95.2	33.5	31.8	39.3	88.7	29.0	36.0	41.3	86.0
DriveAdapter Jia et al. (2023a)	29.5	40.0	40.3	92.4	27.6	41.8	40.0	95.3	34.0	34.7	39.0	89.0	30.8	40.3	43.7	84.5
DriveTrans. Jia et al. (2025)	31.0	40.3	43.1	91.9	28.9	40.2	42.7	94.5	36.6	36.2	43.1	88.5	31.3	47.0	45.0	85.8
ORION Fu et al. (2025)	37.4	65.3	54.8	93.5	34.4	62.9	51.4	95.5	43.8	70.1	61.3	91.4	40.5	68.7	59.5	87.8
Diffu.Drive Liao et al. (2025)	34.2	58.3	49.1	91.4	30.8	56.0	45.3	93.4	41.3	64.0	56.4	89.4	38.1	59.6	54.8	85.3
SimLingo Renz et al. (2025)	42.3	70.5	63.3	94.8	39.1	66.7	60.9	96.3	50.4	77.4	70.7	92.1	43.8	76.2	62.1	92.7
KnowVal Xia et al. (2025a)	46.6	73.8	69.3	97.4	42.6	70.5	66.0	98.3	56.8	79.9	78.5	96.3	47.9	79.0	69.4	95.3

- **SimLingo** Renz et al. (2025): a language-guided E2E driving policy that aligns natural-language prompts with driving action sequences for interpretable human-in-the-loop control.
- **KnowVal** Xia et al. (2025a): a knowledge-augmented and value-guided driving system that retrieves open-world knowledge and evaluates candidate trajectories with a learned value model.

All methods are trained on the Bench2Drive training data. For closed-loop evaluation, all methods are run on the 330 test routes described in Section 3, and metrics are computed accordingly.

4.2 Results

Table 3 summarizes the overall performance of the completed runs on HiDrive, together with split-level results on the Basic, Hard, and Thorny ability sets. The results show that conventional end-to-end autonomous driving methods remain strongly limited on HiDrive: even the best, DiffusionDrive, achieves only 34.2 Overall DS. This indicates that, in a closed-loop environment involving long-tail objects, complex interactions, and legal and ethical constraints, methods that mainly rely on perception-prediction-planning pipelines or trajectory-generation paradigms still struggle to generalize to high-level driving capability evaluation.

In contrast, VLA and language-enhanced paradigms exhibit greater closed-loop robustness and generalization. ORION and SimLingo both outperform the conventional end-to-end baselines, while KnowVal achieves the best overall performance with $DS = 46.6$. Compared with SimLingo, KnowVal further improves DS by 4.3 points and LS by 6.0 points. These results suggest that injecting external or structured knowledge and using value-guided evaluation can help models make more stable decisions while adhering to legal compliance, risk avoidance, and ethical constraints.

Training on HiDrive. HiDrive is built on CARLA and natively supports reinforcement learning training, as well as training with expert models such as CARLA-Roach. The results above evaluate Bench2Drive-trained models under zero-shot transfer. We further fine-tune SimLingo on HiDrive and train a Roach reinforcement-learning expert in the same environment, and report the resulting driving scores in Table 4.

Table 4: **Training results on HiDrive.** We report driving scores for Bench2Drive-trained SimLingo, HiDrive-fine-tuned SimLingo, and the Roach expert across Overall and the Basic/Hard/Thorny sets. All values are percentages (%).

Method	**Overall**	**Basic**	**Hard**	**Thorny**
SimLingo trained on Bench2Drive	42.3	39.1	50.4	43.8
SimLingo fine-tuned on HiDrive	45.4	43.8	52.9	49.0
Roach expert	88.0	89.2	93.5	77.4

Fine-tuning SimLingo on HiDrive raises overall DS from 42.3 to 45.4. The Basic, Hard, and Thorny scores improve from 39.1/50.4/43.8 to 43.8/52.9/49.0, respectively, with the largest gain on the Thorny Set. The Roach expert reaches 88.0 overall, confirming that HiDrive supports in-domain policy learning. Its score on the Thorny Set is only 77.4, reflecting that the scenarios in this set require models to execute relatively complex driving strategies.

One seemingly counterintuitive observation is that the DS or RC on the Hard and Thorny sets is often higher than that on the Basic set. This does not imply that Hard or Thorny scenarios are inherently easier. Instead, it is related to the ability taxonomy and metric definitions in HiDrive. The Basic Set contains many tasks that are highly sensitive to low-level control stability and local interactions, in which collisions, deadlocks, or route interruptions can easily reduce RC and DS. By contrast, some Hard and Thorny scenarios emphasize high-level decision-making. For example, when passing through a puddle, a vehicle may continue driving even if it does not fully comply with the intended ethical behavior, so the route is less likely to be interrupted. This can lead to a higher RC and consequently a higher DS. However, Hard and Thorny scenarios are not easier: for most methods, ES decreases from the Basic set to the Hard and Thorny sets, indicating that legal and ethical capabilities remain challenging.

Fine-grained ability analysis. To complement the split-level comparison above, Table 5 reports representative ability-level scores spanning foundational operational skills and higher-order decision-making capabilities. Together, these results show that HiDrive supports fine-grained diagnosis beyond aggregate Basic, Hard, and Thorny scores. In particular, the low scores on value-priority dilemmas indicate that norm-sensitive reasoning remains difficult even for the strongest aggregate baselines.

Table 5: **Selected ability-level scores.** We report six representative abilities across ten evaluated methods. All values are percentages (%).

Ability	TCP	UniAD-Base	VAD	ThinkTwice	DriveAdapter	DriveTrans.	ORION	Diffu.Drive	SimLingo	KnowVal
Emergency avoidance	26.45	34.60	28.57	33.69	39.22	36.50	42.12	48.99	54.10	64.38
Obstacle detouring	17.11	22.37	19.75	29.38	23.50	24.91	33.81	24.52	38.23	40.79
Signalized turning	18.71	21.52	20.59	28.77	35.23	30.68	37.28	29.73	44.38	55.52
Partial sensor blindness	26.08	29.96	25.65	41.74	37.01	35.43	48.68	41.08	52.54	68.13
Value-priority dilemma	5.80	8.10	9.89	12.16	14.85	11.41	17.53	17.85	16.64	15.55
Unknown-object distancing	21.18	26.66	30.08	29.18	41.26	26.36	39.80	48.84	56.12	50.75

Overall, HiDrive reveals clear differences among current autonomous driving paradigms: conventional end-to-end methods still show substantial limitations in long-tail closed-loop generalization, legal compliance, and social-norm understanding, while VLA-style methods benefit from stronger semantic understanding and instruction alignment. Furthermore, the experimental results of KnowVal demonstrate that external knowledge injection and value-guided reasoning can further improve legal, ethical, and generalization capabilities.

5 Conclusion

In this paper, we propose HiDrive, a new closed-loop benchmark that systematically incorporates diverse rare objects and a richer set of driving capabilities, including rule compliance, moral reasoning, and emergency response. By extending evaluation beyond collision avoidance to encompass collision and braking behavior, traffic-rule compliance, and moral-reasoning indicators, HiDrive provides a more holistic assessment of driving intelligence. Our experiments reveal that state-of-the-art planners, which have already achieved near-saturated scores on widely used benchmarks, still exhibit significant performance degradation on HiDrive, particularly in long-tail perception and norm-sensitive scenarios. This highlights that current models, while highly competent in routine situations, are far from robust in the complex, ambiguous, and interactive conditions that characterize real-world driving. Built on a more advanced physics-rendering engine, HiDrive offers high-fidelity visual realism and physically plausible dynamics, making it a challenging and diagnostically valuable testbed for future research. We believe HiDrive will encourage the community to move beyond chasing incremental gains on saturated benchmarks and instead focus on building autonomous driving systems that are genuinely safe, socially aware, and prepared for the rare but critical events of everyday traffic.

AI USE STATEMENT

We used generative AI tools to polish the text and check grammar. All AI-assisted outputs were reviewed and verified by the authors. The authors reviewed all AI-assisted changes and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The HiDrive software, source code, digital assets, and documentation are available in this anonymous repository: https://huggingface.co/HiDriveBenchmark/HiDriveBenchmark.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 3

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *CoRL*, 2025. 1, 3

Canyu Chen, Yuguang Yang, Zhewen Tan, Yizhi Wang, Ruiyi Zhan, Haiyan Liu, Xuanyao Mao, Jason Bao, Xinyue Tang, Linlin Yang, Bingchuan Sun, Yan Wang, and Baochang Zhang. Devil is in narrow policy: Unleashing exploration in driving VLA models. *arXiv preprint arXiv:2603.06049*, 2026. 1, 4

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *NeurIPS*, 2024. 1, 3

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017. 3, 4, 5

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. ORION: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In *ICCV*, 2025. 3, 7, 8

Xinmeng Hou, Wuqi Wang, Long Yang, Hao Lin, Jinglun Feng, Haigen Min, and Xiangmo Zhao. DriveAgent: Multi-agent structured reasoning with LLM and multimodal sensor fusion for autonomous driving. *IEEE RAL*, 2025. 3

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *ECCV*, 2022. 3

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, 2023. 3, 7, 8

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 *ICCV*, 2023a. 3, 7, 8

Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *CVPR*, 2023b. 3, 7, 8

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *NeurIPS*, 2024. 1, 3, 4, 5

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. DriveTransformer: Unified transformer for scalable end-to-end autonomous driving. In *ICLR*, 2025. 3, 7, 8

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 *ICCV*, 2023. 3, 7, 8

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 3

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. DiffusionDrive: Truncated diffusion model for end-to-end autonomous driving. In *CVPR*, 2025. 3, 7, 8

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving. In *ECCV*, 2024. 3

Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. SimLingo: Vision-only closed-loop autonomous driving with language-action alignment. In *CVPR*, 2025. 4, 8

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *ECCV*, 2024. 3

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. SparseDrive: End-to-end autonomous driving via sparse scene representation. In *ICRA*, 2025. 3

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. SparseDriveV2: Scoring is all you need for end-to-end autonomous driving. *arXiv preprint arXiv:2603.29163*, 2026. 1

Linbo Wang, Yupeng Zheng, Qiang Chen, Shiwei Li, Yichen Zhang, Zebin Xing, Qichao Zhang, Xiang Li, Deheng Qian, Pengxuan Yang, Yihang Dong, Ce Hao, Xiaoqing Ye, Junyu Han, Yifeng Pan, and Dongbin Zhao. Latent-WAM: Latent world action modeling for end-to-end autonomous driving. *arXiv preprint arXiv:2603.24581*, 2026. 1, 4

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. PARA-Drive: Parallelized architecture for real-time autonomous driving. In *CVPR*, 2024. 3

Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *NeurIPS*, 2022. 3, 7, 8

Zhongyu Xia, Wenhao Chen, Yongtao Wang, and Ming-Hsuan Yang. KnowVal: A knowledge-augmented and value-guided autonomous driving system. *arXiv preprint arXiv:2512.20299*, 2025a. 1, 4, 5, 8

Zhongyu Xia, Zhiwei Lin, Yongtao Wang, and Ming-Hsuan Yang. HENet++: Hybrid encoding and multi-task learning for 3d perception and end-to-end autonomous driving. *arXiv preprint arXiv:2511.07106*, 2025b. 3

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE RAL*, 2024. 3

Huaiyuan Yao, Longchao Da, Vishnu Nandam, Justin Turnau, Zhiwei Liu, Linsey Pang, and Hua Wei. CoMAL: Collaborative multi-agent large language models for mixed-autonomy traffic. In *SDM*, 2025. 3

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. GenAD: Generative end-to-end autonomous driving. In *ECCV*, 2024. 3

A Additional Analyses

A.1 Penalty-Coefficient Design and Sensitivity

Selection principles. We choose the coefficients in three stages. First, collision coefficients follow the CARLA Leaderboard and Bench2Drive-style protocols where possible, preserving comparability with established closed-loop evaluation. Second, law-related events are ordered by violation severity: events that clearly violate mandatory rules or invalidate the intended behavior receive stronger penalties, while context-dependent violations receive milder or rule-based treatment. Third, for norm-sensitive events without established benchmark coefficients, 14 participants ranked the relative severity of the scenarios. We use this survey-derived partial order, align it with the scale of the law-related penalties, and assign coefficients that preserve the resulting severity ordering. Scenario-specific overrides are used only when applying the generic event penalty would contradict the intended behavior of a dilemma scenario.

Table 6: **Penalty coefficients in HiDrive.** The table lists representative update coefficients for legal score and ethics score. Some collision rules use brake-aware relief, and dilemma scenarios may use explicit overrides.

Event type	Coefficient	Applied to
COLLISION_PEDESTRIAN	0.5 / 0.6	LS
COLLISION_VEHICLE	0.6 / 0.72	LS
COLLISION_STATIC	0.65 / 0.78	LS
SCENARIO_TIMEOUT	0.7	LS
TRAFFIC_LIGHT_INFRACTION	0.7	LS
STOP_INFRACTION	0.8	LS
POLICE_STOP_VIOLATION	0.0	LS
OUTSIDE_ROUTE_LANES_INFRACTION	percentage-based / percentage-based	LS
MIN_SPEED_INFRACTION	rule-based (no fixed scalar)	LS (via speed legality)
BrakeFailureDilemma + COLLISION_VEHICLE	0.85 (override)	LS
RedLightEmergencyYield + TRAFFIC_LIGHT_INFRACTION	1.0 (override)	LS
...	...	...
YIELD_TO_EMERGENCY_VEHICLE	0.5	ES
PUDDLE_ETHICS_INFRACTION	0.8	ES
DOOR_PASS_SPEED_ETHICS_INFRACTION	0.8	ES
UNSAFE_ROADSIDE_MERGE_ETHICS_INFRACTION	0.7	ES
WEAVE_CLOSE_DISTANCE_ETHICS_INFRACTION	0.7	ES
SLOW_LEAD_NO_OVERTAKE_ETHICS_INFRACTION	0.7	ES
SPEED_BUMP_OVERSPEED_ETHICS_INFRACTION	0.8	ES
...	...	...

Brake-aware relief. In urgent scenarios, a collision can be unavoidable even when the policy reacts appropriately. We therefore distinguish complete hazard neglect from an active mitigation attempt. For entries written as "a/b" in Table 6, a is the default coefficient and b is used only when braking is detected immediately before the collision. Let

$$\mathcal{E}_{\text{relief}} = \{\texttt{COLLISION_PEDESTRIAN}, \texttt{COLLISION_VEHICLE}, \texttt{COLLISION_STATIC}\}.$$

If event $e \in \mathcal{E}_{\text{relief}}$ occurs at frame f_e, relief is triggered when

$$\text{BrakeNear}(f_e) = 1 \iff \exists\, t \in [f_e - 2, f_e],\ \max(\texttt{brake}_t, \mathbf{1}\{\texttt{hand_brake}_t\}) \geq 0.2.$$

This binary rule avoids rewarding increasingly aggressive braking, which could introduce discomfort or secondary closed-loop risks. Entries without a slash always use the same coefficient.

Fixed-coefficient sensitivity. To test whether the conclusions depend on the selected absolute magnitudes, we replace every penalty coefficient with 0.5 and recompute all driving scores. Table 7

shows that the overall ranking trend remains consistent with the designed coefficients. In particular, language-enhanced methods remain stronger than conventional end-to-end baselines, and KnowVal remains the best aggregate method. The benchmark conclusions are therefore not driven solely by coefficient tuning.

Table 7: **Fixed-coefficient sensitivity.** We report driving scores when every penalty coefficient is fixed to 0.5. All values are percentages (%).

Set	TCP	UniAD-Base	VAD	ThinkTwice	DriveAdapter	DriveTrans.	ORION	Diffu.Drive	SimLingo	KnowVal
Basic	16.0	19.9	16.8	24.8	26.1	28.1	32.5	29.0	36.6	39.8
Hard	16.2	18.2	16.0	22.2	23.1	23.9	29.3	27.1	33.5	35.3
Thorny	15.6	21.3	18.9	30.2	28.8	31.1	38.9	34.8	45.0	47.0
Full HiDrive	17.3	19.0	16.5	23.9	28.0	27.8	36.7	32.3	38.6	42.8

A.2 Event Acceptance and Scenario Scalability

High-level dilemma scenarios do not require one predefined trajectory. The evaluator checks closed-loop safety, legal, and norm-sensitive constraints, so multiple reasonable behaviors are accepted if they avoid the corresponding infractions. For example, in the puddle scenario, either changing lanes or slowing below the triggering threshold is accepted. Scenario templates can also be inserted into compatible roads and intersections using lane-level annotations, allowing new routes to be generated without rebuilding each event from scratch.

The brake-aware relief test is deliberately binary: it distinguishes ignoring a hazard from making a meaningful mitigation attempt. The score is not monotonic in braking intensity because that could reward unnecessarily aggressive emergency braking and introduce discomfort or secondary closed-loop risks.