PLANCONTEXT: BENCHMARKING IN-CONTEXT DRIVING DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world driving requires context-aware decision-making. Existing end-to-end driving benchmarks primarily assess trajectory quality through safety, compliance, comfort, and driving progress, leaving context-dependent decision-making insufficiently characterized. To fill this gap, we present **PlanContext**, a benchmark for evaluating **in-context driving decision-making**. PlanContext constructs paired evaluations by varying high-level driving intentions while keeping the observed scene fixed, enabling a natural and controllable assessment of context-dependent decision-making. We further define six representative driving contexts and jointly evaluate contextual decision correctness, road compliance, and dynamic safety with class-balanced metrics. Experiments on 9,561 scene–condition instances from NAVSIM and 23 public end-to-end driving models show that high conventional driving-quality scores do not consistently correspond to strong route-level in-context driving decision-making. Further analysis shows that model behavior varies with decision timing, road structure, and traffic interaction depending on the decision required by the current context. Our results highlight in-context driving decision-making as an important complementary dimension for evaluating end-to-end driving systems.

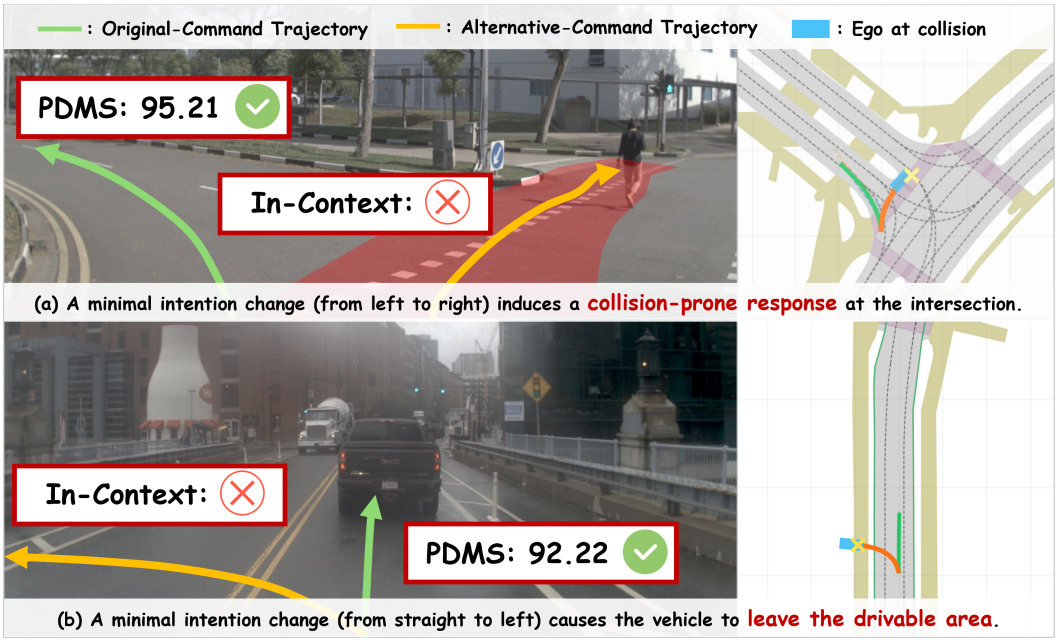


Figure 1: **PlanContext.** Conventional driving metrics primarily evaluate trajectory quality in terms of safety, compliance, comfort, and driving progress. However, models with high PDMS can still produce context-inappropriate trajectories when navigation intentions change. PlanContext uses a natural and controllable navigation intervention to evaluate in-context driving decision-making.

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.

1 INTRODUCTION

Autonomous driving has attracted growing attention from both academia and industry due to its potential to improve road safety and transportation efficiency (Chen et al., 2024; Janai et al., 2020). The community has developed a wide range of end-to-end driving models that integrate sensory observations, ego states, and high-level driving intentions to directly predict future trajectories or driving actions (Chitta et al., 2022; Hu et al., 2023; Zhou et al., 2026a). As model families and technical paradigms continue to diversify, reliable evaluation and fair comparison of driving performance have become increasingly important for measuring progress in the field.

However, comprehensively characterizing the decision-making capabilities of end-to-end driving models remains challenging. Existing benchmarks primarily assess trajectory quality through metrics such as collision avoidance, road compliance, comfort, and driving progress (Caesar et al., 2020; Jia et al., 2024; Cao et al., 2025; Zhang et al., 2026). While these metrics are effective for measuring whether a model produces safe and reasonable trajectories, a high-quality trajectory alone does not establish that the model has correctly interpreted the current driving context and made an appropriate decision. As shown in Fig. 1, models with high PDMS under the original navigation command can still respond inappropriately when the command changes. In Fig. 1(a), the vehicle has already started a left-turn maneuver, and a newly issued right-turn command should not trigger an immediate reversal that drives the vehicle toward a pedestrian. Similarly, in Fig. 1(b), when a turning instruction is given before the corresponding maneuver becomes locally feasible, the policy should continue to follow the current road rather than turn immediately and leave the drivable area. These cases highlight that driving decisions should be made in accordance with the current driving context. A reliable autonomous driving policy should jointly consider contextual information such as the current vehicle state, road geometry, and high-level navigation intention to make an appropriate decision. Yet existing benchmarks offer limited insight into such in-context decision-making capability, allowing context-inappropriate decision failures to remain hidden behind strong trajectory-level performance.

To address this gap, we introduce **PlanContext**, a benchmark for systematically evaluating **in-context driving decision-making** in end-to-end autonomous driving. PlanContext keeps sensor observations and vehicle history fixed while applying controlled changes to high-level navigation intentions, enabling paired comparison of model decisions under the same environmental context. Based on road topology, target reachability, decision timing, and ongoing vehicle motion, we construct representative driving contexts and define context-specific requirements for whether an intention should be executed, deferred, rejected, or preserved. PlanContext jointly evaluates contextual decision correctness, road compliance, and dynamic safety, and further analyzes model behavior under different contextual factors. In this way, PlanContext extends end-to-end driving evaluation beyond trajectory quality toward a systematic assessment of intention–context–decision consistency.

Based on PlanContext, we construct 9,561 strictly filtered scene–condition instances from NAVSIM (Dauner et al., 2024) and evaluate 23 publicly available end-to-end driving models spanning three paradigms: vision/state-to-action, vision-language-action, and world-model-based action. Model rankings under conventional driving-quality evaluation differ substantially from those under route-level in-context evaluation. Across all 23 models, contextual success reaches only 22.46% at best when the driving context requires executing an immediately actionable alternative route. Moreover, no consistent performance trend emerges across model paradigms: VA, VLA, and WA models all exhibit substantial within-paradigm variation in in-context decision-making performance. Our key contributions are threefold:

- **We introduce PlanContext, a benchmark for in-context driving decision-making with 9,561 curated scene–condition instances.** It enables systematic evaluation of whether end-to-end models make appropriate decisions by jointly considering driving context and high-level intentions.
- **PlanContext provides evaluation criteria and metrics for in-context driving decision-making across six scenario categories.** The evaluation jointly assesses contextual decision correctness, road compliance, and dynamic safety, with separate metrics for route-level and lane-level performance.
- **Our evaluation of 23 end-to-end driving models shows that conventional driving-quality metrics do not fully capture in-context decision-making capability.** Further analyses reveal category-dependent performance patterns across decision timing, road structure, and traffic interactions.

2 Related Work

End-to-End Autonomous Driving. End-to-end autonomous driving directly maps sensory observations and vehicle states to future trajectories or actions. Recent approaches span perception–planning architectures, direct trajectory regression, generative planners, vision–language–action models, and world-model-based driving systems (Jiang et al., 2023; Chitta et al., 2022; Liao et al., 2025; Zhou et al., 2026b; Wang et al., 2026a). Beyond sensory inputs, modern models increasingly incorporate navigation goals and structured scene information into decision-making. Recent work further shows that decision-relevant traffic elements, including traffic lights, road signs, and lane topology, consistently affect planning performance across diverse end-to-end paradigms (Zhang et al., 2026). These developments motivate evaluation beyond aggregate trajectory quality toward how models use contextual information when making driving decisions.

Evaluation of End-to-End Driving. End-to-end driving evaluation has evolved from open-loop trajectory similarity toward more comprehensive measurements of driving quality and robustness. NAVSIM (Dauner et al., 2024) introduces data-driven non-reactive simulation and evaluates predicted trajectories through collision, road compliance, comfort, time-to-collision, and driving progress. Closed-loop benchmarks such as Bench2Drive (Jia et al., 2024) further evaluate interactive driving performance, while NAVSIMv2 (Cao et al., 2025) augments logged data with synthetic observations to assess recovery from ego deviations. Fail2Drive (Gerstenecker et al., 2026) further examines closed-loop generalization by pairing distribution-shifted routes with in-distribution counterparts under changes in appearance, layout, and agent behavior. These benchmarks substantially improve the evaluation of trajectory quality, safety, and robustness. PlanContext provides a complementary evaluation perspective by examining whether driving decisions remain appropriate under different combinations of high-level intentions and environmental contexts, thereby revealing aspects of decision-making that are not captured by conventional driving-quality metrics.

Instruction-Conditioned and Intervention-Based Evaluation. High-level navigation commands help resolve behavioral ambiguity in end-to-end driving (Codevilla et al., 2018; Chitta et al., 2022). NavControl (Wu et al., 2025) constructs feasible alternative commands in NAVSIM and evaluates navigation compliance alongside driving quality. ICR-Drive (Hamid et al., 2026) assesses robustness to goal-preserving and goal-conflicting language perturbations on matched CARLA routes, using the original navigation objective as the evaluation reference. Beyond driving policies, What-If World (Cai et al., 2026) evaluates whether video world models produce the expected physical outcomes under paired prompts differing in a single physical variable. PlanContext targets context-dependent decision correctness through paired intention changes with observations and vehicle history fixed. Its criteria account for road topology, target reachability, decision timing, and ongoing motion, assessing both the semantic validity of individual predictions and whether paired predictions exhibit context-appropriate behavioral adjustment or preservation.

3 PlanContext: Evaluating In-Context Driving Decision-Making

PlanContext evaluates in-context driving decision-making through paired navigation interventions and context-aware assessment. We first introduce the intervention protocol and the construction of representative driving contexts in Sections 3.1 and 3.2. Section 3.3 then defines the contextual decision evaluation criteria, followed by the class-balanced performance metrics in Section 3.4.

3.1 Paired Navigation Intervention

Given a driving scene s_i, its original navigation intention c_i^N, and an alternative intention c_i^F, an evaluated model f_θ predicts two future trajectories:

$$T_i^N = f_\theta(s_i, c_i^N), \qquad T_i^F = f_\theta(s_i, c_i^F). \tag{1}$$

Each trajectory contains eight future ego poses over a 4-second horizon. PlanContext keeps the sensor observations and vehicle history in s_i unchanged and only modifies the high-level navigation intention, yielding paired predictions under the same environmental context.

For each condition $x \in \{N, F\}$, let $\mathcal{A}(s_i, c_i^x)$ denote the set of behaviors that are semantically valid under scene s_i and intention c_i^x. The semantic-validity indicator is defined as

$$S_i^x = \mathbb{1}\left[T_i^x \in \mathcal{A}(s_i, c_i^x)\right], \qquad x \in \{N, F\}. \tag{2}$$

This formulation allows multiple valid future behaviors under the same intention and avoids treating a single recorded trajectory as the unique correct outcome.

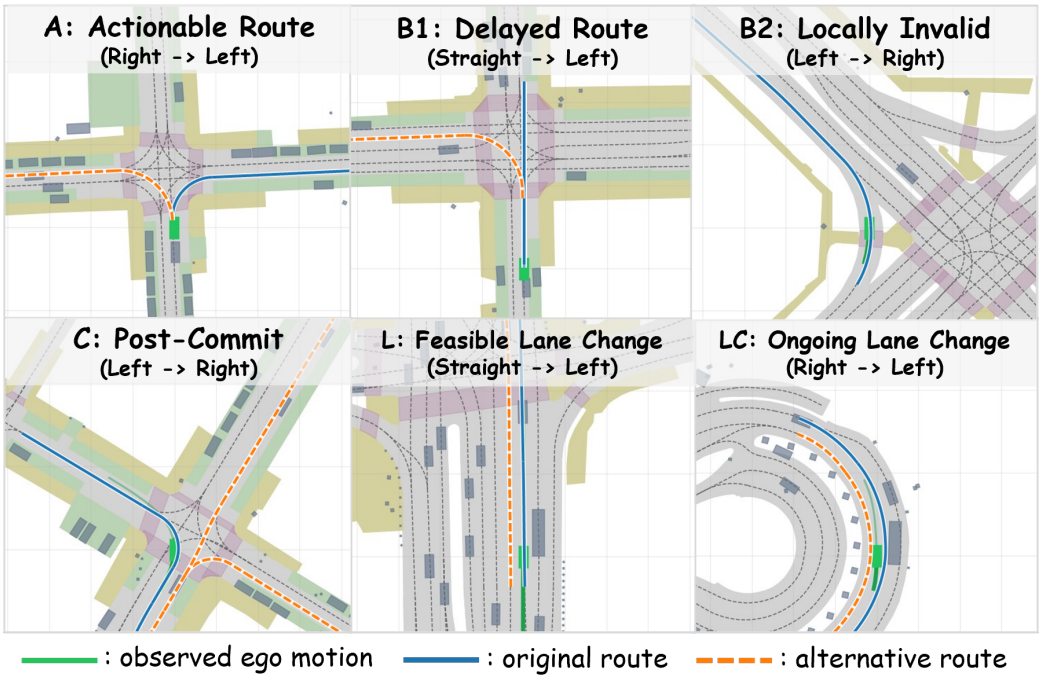


Figure 2: Representative examples of the six PlanContext categories. Parentheses indicate the change in navigation intention.

The logged trajectory T_i^{GT} is used to characterize the natural road evolution, maneuver stage, and route-commitment state of the scene. These cues support the construction of the context-dependent valid behavior sets $\mathcal{A}(s_i, c_i^N)$ and $\mathcal{A}(s_i, c_i^F)$. The paired relationship between T_i^N and T_i^F is evaluated separately in Section 3.3.

3.2 Context-Aware Scenario Construction

The valid behavior sets introduced in Section 3.1 are determined by the driving context of each scene–condition instance. We characterize this context using road topology, target reachability, decision timing, and the ongoing motion state of the ego vehicle. Together, these factors determine how a high-level driving intention can be appropriately realized in the current scene.

PlanContext organizes the resulting instances into four contextual requirements: *follow*, *wait*, *reject*, and *preserve*. *Follow* applies when the alternative target is currently reachable and should be executed within the evaluation horizon. *Wait* describes cases where the target becomes feasible in the near future while the current state still favors maintaining the existing route. *Reject* covers locally unreachable targets or cases where the vehicle has already passed the appropriate decision point. *Preserve* applies when an ongoing maneuver should remain temporally consistent after the intention changes. We instantiate these requirements with six complementary scenario categories, summarized in Table 1 and illustrated in Fig. 2.

Table 1: PlanContext scenario categories and contextual requirements.

Category	Context condition	Requirement
A	Target route reachable within 2 s	Follow
B1	Target route reachable in 4.5–8 s	Wait
B2	No locally feasible target	Reject
C	Ego beyond route commitment point	Reject
L	Adjacent target lane geometrically and topologically reachable	Follow
LC	Lateral maneuver already in progress	Preserve

The route-level categories A, B1, B2, and C capture different stages of navigation feasibility and route commitment. A represents an immediately actionable route change, while B1 contains targets that become feasible beyond the current decision window. B2 captures locally unsupported intentions, and C describes scenes in which the ego vehicle has already committed to the original route. The lane-level categories complement these route-level contexts: L evaluates feasible adjacent-lane transitions, while LC evaluates the continuity of an ongoing lateral maneuver.

For A, B1, B2, C, and L, each category defines the semantic behavior sets $\mathcal{A}(s_i, c_i^N)$ and $\mathcal{A}(s_i, c_i^F)$ used to compute S_i^N and S_i^F in Section 3.1. LC instead defines a continuity requirement for the

ongoing lateral maneuver, as detailed in Section 3.3. We retain only scene–condition instances with sufficiently clear contextual evidence for reliable category assignment. Detailed scenario mining, filtering, and category assignment criteria are provided in Sec. A.

To further assess benchmark validity, three reviewers with backgrounds in autonomous driving independently inspected a randomly sampled subset of 500 scene–condition instances and 11,500 model-instance evaluations from all 23 evaluated models after the evaluation rules were finalized. The reviewers were blinded to the automated evaluation results during independent assessment. Following discussion, approximately 0.7% of the inspected model-instance evaluations were judged inconsistent with the automated evaluation outcomes. Detailed findings are provided in Appendix D.9.

3.3 Contextual Decision Evaluation

Given the contextual requirement assigned in Section 3.2, PlanContext evaluates each scene–condition instance through three components: contextual decision correctness, road compliance, and dynamic safety.

Paired behavioral correctness. We first characterize how the prediction changes when the navigation intention changes. The original and alternative trajectories are projected into a road-relative coordinate system, where longitudinal progress, lateral displacement, and heading deviation are measured with respect to the local road geometry. The trajectories are aligned by road progress to reduce the influence of speed differences on behavioral comparison.

Let $k_i \in \{A, B1, B2, C, L, LC\}$ denote the PlanContext category of instance i. We define a category-specific paired behavior function

$$P_i = \Phi_{k_i}\left(T_i^N, T_i^F, s_i\right), \tag{3}$$

where Φ_{k_i} evaluates whether the behavioral change between the original and alternative predictions satisfies the requirement of category k_i.

For the *follow* categories A and L, P_i requires an identifiable path change toward the alternative target. For B1, B2, and C, it requires the alternative intention to avoid inducing an inappropriate target-directed deviation. LC is treated separately because its objective is to preserve the continuity of an ongoing lateral maneuver rather than to evaluate two independent semantic outcomes. We therefore assign LC predictions a continuity status $P_i^{LC} \in \{\text{correct}, \text{incorrect}, \text{uncertain}\}$. Detailed category-specific behavioral criteria and trajectory comparison rules are provided in Sec. B.

For A, B1, B2, C, and L, contextual decision correctness combines the semantic validity of both predictions with their paired behavioral relationship:

$$C_i = \begin{cases} S_i^N \wedge S_i^F \wedge P_i, & k_i \in \{A, B1, B2, C, L\}, \\ 1, & k_i = LC,\ P_i^{LC} = \text{correct}, \\ 0, & k_i = LC,\ P_i^{LC} = \text{incorrect}. \end{cases} \tag{4}$$

LC instances with $P_i^{LC} = \text{uncertain}$ are excluded from subsequent success-rate aggregation.

Road compliance. Road compliance is evaluated using the full ego-vehicle footprint. Let $\mathcal{B}(T_{i,t}^F)$ denote the vehicle footprint at future step t and Ω_i^F the road region permitted by the corresponding context. The road-compliance indicator is

$$R_i^F = \prod_{t=1}^{8} \mathbb{1}\left[\mathcal{B}(T_{i,t}^F) \subseteq \Omega_i^F\right]. \tag{5}$$

Dynamic safety. Dynamic safety is evaluated using the NAVSIM safety criteria, including collision, drivable-area compliance, wrong-way driving, traffic-light violations, time-to-collision, and comfort. We denote the resulting safety indicator as E_i^F.

The final contextual success for instance i is defined as

$$q_i = C_i \wedge R_i^F \wedge E_i^F. \tag{6}$$

Thus, a successful instance requires a context-consistent decision response, road-compliant motion, and dynamic safety. For A, B1, B2, C, and L, contextual correctness additionally requires semantic validity under both navigation conditions.

3.4 CLASS-BALANCED PLANCONTEXT METRICS

The six PlanContext categories are highly imbalanced in sample size, making a direct micro-average sensitive to the empirical category distribution. We therefore first compute the contextual success rate independently for each category:

$$S_k = \frac{1}{|\mathcal{D}_k|} \sum_{i \in \mathcal{D}_k} q_i, \qquad k \in \{A, B1, B2, C, L, LC\}, \tag{7}$$

where $\mathcal{D}_k$ denotes the set of valid scene–condition instances used for metric aggregation in category k. LC instances labeled as uncertain are excluded.

We aggregate the scores of A, B1, B2, and C to evaluate route-level in-context driving decision-making. The scores of L and LC are reported separately to evaluate lane-level in-context driving decision-making.

The Navigation Action Score (NAS) is the contextual success rate in category A, where the alternative route is immediately actionable. The Context Preservation Score (CPS) measures contextual success when the alternative navigation intention should be deferred or rejected, averaged across categories B1, B2, and C:

$$\mathrm{CPS} = \frac{S_{B1} + S_{B2} + S_C}{3}. \tag{8}$$

We define the Navigation Context Score (NCS) as the geometric mean of NAS and CPS to summarize route-level in-context driving decision-making:

$$\mathrm{NCS} = \sqrt{\mathrm{NAS} \cdot \mathrm{CPS}}. \tag{9}$$

The geometric mean penalizes imbalanced performance across NAS and CPS.

For lane-level evaluation, we report the Lane Adaptation Score (LAS) and Maneuver Continuity Score (MCS), defined as the contextual success rates in categories L and LC, respectively. We report all five metrics together with the six category-wise scores.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmark. PlanContext is constructed from the NAVSIM `navtest` split. After scenario mining and context-aware filtering, the benchmark contains 9,561 scene–condition instances from 5,911 unique physical scenes. The six contextual categories include 334 A, 251 B1, 5,285 B2, 580 C, 2,716 L, and 395 LC instances. For each retained instance, the sensor observations and vehicle history are fixed, while the evaluated model is queried under both the original navigation intention and the corresponding alternative intention.

Models. We evaluate 23 publicly available end-to-end driving models covering three representative paradigms. Vision/state-to-action (VA) models directly predict driving trajectories from sensory or state inputs, vision-language-action (VLA) models incorporate vision-language representations into planning, and world-model-based action (WA) models explicitly predict or encode future scene dynamics. All models are evaluated using publicly released or official assets, with navigation conditions explicitly involved in inference. Their outputs are standardized as eight future ego poses over a 4-second horizon.

Evaluation Protocol. For each model, we report its conventional driving performance on the full NAVSIM `navtest` split using PDMS. To enable a scene-matched comparison with PlanContext, we additionally compute matched PDMS (mPDMS) under the original navigation condition on the same 5,911 physical scenes selected by PlanContext. Each physical scene contributes once to mPDMS, even when it is associated with multiple alternative navigation conditions. PlanContext then performs paired inference under the original and alternative navigation intentions on the 9,561 scene–condition instances. We report NAS, CPS, and NCS for route-level in-context driving decision-making, and LAS and MCS for lane-level evaluation. We additionally report contextual success rates for all six categories.

Table 2: Overall comparison of 23 end-to-end driving models on PlanContext.

Model	Venue	PDMS	mPDMS	NAS	CPS	NCS	LAS	MCS
VA Models								
DrivoR (Kirby et al., 2026)	CVPR 2026	93.96	83.56	19.76	61.23	34.78	0.00	92.62
GoalFlow (Xing et al., 2025)	CVPR 2025	88.31	79.82	15.57	63.89	31.54	0.07	87.82
TransFuser (Chitta et al., 2022)	IEEE TPAMI	83.39	74.16	13.77	52.98	27.01	0.15	73.59
SparseDriveV2 (Sun et al., 2026)	ECCV 2026	92.22	81.22	12.57	56.25	26.60	0.15	93.64
DiffusionDrive (Liao et al., 2025)	CVPR 2025	88.05	79.80	3.29	65.44	14.68	0.00	92.11
StateMLP (Dauner et al., 2024)	NeurIPS 2024	65.52	56.95	15.57	8.68	11.63	**8.73**	31.55
CLOVER (Ang et al., 2026)	Arxiv 2026	94.48	94.35	0.60	57.35	5.86	0.22	85.82
DiffusionDriveV2 (Zou et al., 2025)	Arxiv 2026	90.95	90.89	0.00	49.51	0.00	0.07	82.19
VLA Models								
OneVL (Lu et al., 2026)	Arxiv 2026	88.92	80.79	**22.46**	68.79	**39.30**	0.00	91.88
ChainFlow-VLA (Wang et al., 2026b)	Arxiv 2026	95.21	83.14	20.96	58.88	35.13	0.04	94.42
ReCogDrive-IL (Xiong et al., 2026)	ICLR 2026	86.00	78.59	21.26	55.39	34.31	0.00	82.56
ReCogDrive-RL (Xiong et al., 2026)	ICLR 2026	90.46	78.93	18.56	53.93	31.64	0.07	84.83
NoRD (Rawal et al., 2026)	CVPR 2026	85.69	75.75	17.66	49.31	29.51	0.18	82.69
AutoVLA (Zhou et al., 2026b)	NeurIPS 2025	92.06	80.92	12.28	59.76	27.08	0.11	86.67
DriveVLA-W0 (Li et al., 2026)	ICLR 2026	87.02	78.54	4.49	63.01	16.82	0.00	91.88
DP-VLA (Zheng et al., 2026)	Arxiv 2026	88.56	88.41	1.80	66.09	10.90	0.00	86.58
Curious-VLA (Chen et al., 2026)	CVPR 2026	89.20	89.32	0.60	60.62	6.02	0.15	87.95
WAM-Flow (Xu et al., 2026)	CVPR 2026	90.27	90.20	0.60	57.27	5.86	0.00	88.27
WA Models								
SUV (Yuan et al., 2026)	Arxiv 2026	90.87	81.63	14.67	**69.94**	32.03	0.00	**94.68**
WoTE (Li et al., 2025)	ICCV 2025	88.22	79.21	13.47	60.36	28.52	0.04	91.88
SimWAM (Zhao et al., 2026)	Arxiv 2026	90.97	90.76	0.30	56.74	4.12	0.00	88.83
Drive-JEPA (Wang et al., 2026a)	Arxiv 2026	93.69	94.18	0.00	52.37	0.00	0.33	85.71
DynVLA Stage 2 (Shang et al., 2026)	ICML 2026	87.20	86.53	0.00	64.54	0.00	0.00	88.61

- PDMS denotes conventional driving quality on the full NAVSIM v1 `navtest` split.
- mPDMS denotes PDMS recomputed under the original navigation condition on the 5,911 physical scenes selected by PlanContext.
- NAS is the contextual success rate in category A.
- CPS is the macro-average of the success rates in categories B1, B2, and C.
- NCS is the geometric mean of NAS and CPS and summarizes route-level in-context driving decision-making.
- LAS and MCS are the contextual success rates in L and LC, respectively, for lane-level evaluation.
- The 95% confidence intervals for NCS are reported in Table 9.

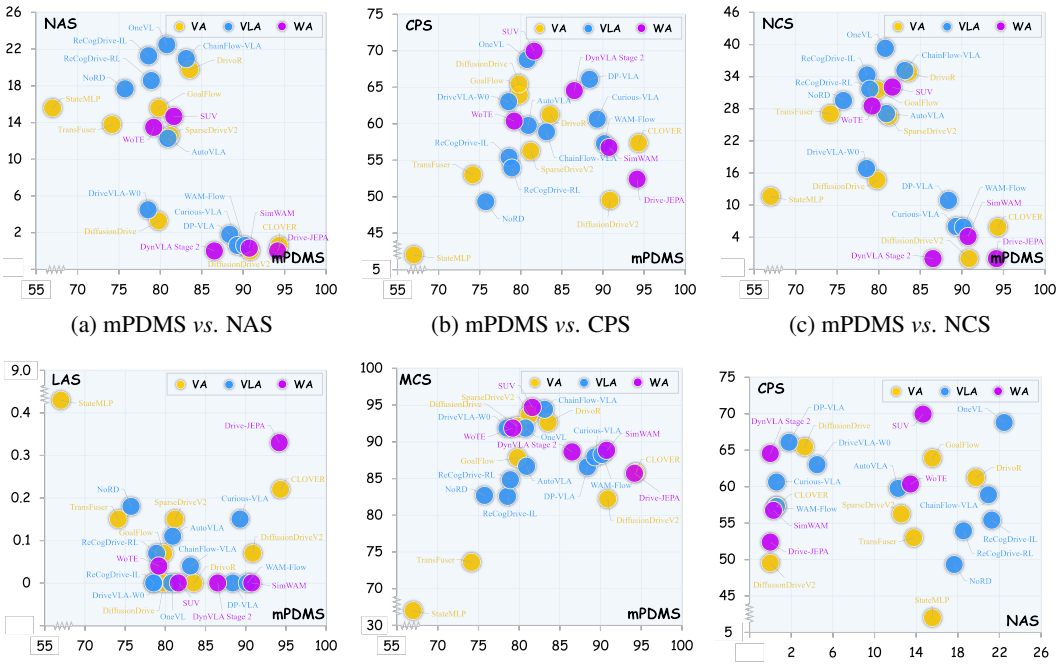
Statistical Analysis. We estimate uncertainty using 10,000 bootstrap resamples over the 5,911 unique scene tokens. All scene–condition instances and model predictions associated with a sampled scene are retained together to preserve the paired evaluation structure. We report percentile-based 95% confidence intervals for NCS. Pairwise model comparisons are computed on matched bootstrap samples, with Benjamini–Hochberg correction applied across all 253 pairwise comparisons.

4.2 Overall Evaluation

Table 2 presents the overall results of the 23 evaluated models. Figures 3a–3e compare conventional driving quality with route-level and lane-level in-context performance. mPDMS evaluates driving quality under the original navigation condition, while NCS evaluates in-context driving decision-making through paired navigation interventions. The Spearman rank correlation between mPDMS and NCS is -0.542, indicating substantial differences in model rankings across the two evaluations. Confidence intervals, pairwise comparisons, and sensitivity analyses of model rankings are provided in Appendices D.3 and D.7.

Figure 3 further reveals two interesting patterns: (1) High mPDMS does not consistently correspond to high NAS. All models with mPDMS above 85 achieve NAS of at most 1.80, while OneVL achieves the highest NAS of 22.46 with an mPDMS of 80.79. (2) Different model architecture paradigms do not exhibit a consistent performance trend on PlanContext. VA, VLA, and WA models all show substantial performance variation within each paradigm, suggesting that the architectural paradigm alone does not consistently explain in-context driving decision-making performance.

4.3 Model Capability Analysis



(a) mPDMS *vs.* NAS (b) mPDMS *vs.* CPS (c) mPDMS *vs.* NCS

(d) mPDMS *vs.* LAS (e) mPDMS *vs.* MCS (f) NAS *vs.* CPS

Figure 3: Driving quality and in-context performance across 23 end-to-end driving models.

Table 3 shows substantial variation in contextual success across the six categories. At the route level, the mean NAS is 10.01, compared with a mean CPS of 57.06. NAS is lower than CPS for 22 of the 23 models, indicating that current models achieve limited success when alternative routes are immediately actionable.

Figure 3f further illustrates differences in model performance across NAS and CPS. DynVLA Stage 2 achieves a CPS of 64.54 but has zero NAS. StateMLP achieves a NAS of 15.57 and a CPS of 8.68.

Table 3: Contextual success across PlanContext categories.

Category	Instances	Success (%)
A	334	10.01
B1	251	62.91
B2	5,285	73.28
C	580	34.98
L	2,716	0.45
LC	395	85.51

Success is averaged over all 23 evaluated models.

At the lane level, Figures 3d and 3e show limited success in executing feasible alternative lane intentions alongside substantially higher success in ongoing-maneuver continuity. LAS is below 0.34 for 22 of the 23 models, including ten with zero scores. StateMLP achieves the highest LAS of 8.73. The consistently low LAS across models identifies the execution of feasible alternative lane intentions as a common failure mode in lane-level in-context driving decision-making.

4.4 Context Factor Analysis

We further investigate how in-context driving performance varies with environmental context under different decision requirements. All analyses use the contextual success indicator q_i defined in Section 3.3 and aggregate predictions from all 23 evaluated models. All comparisons are performed within the same PlanContext category. We group contextual factors into scene-level factors, including decision timing, road geometry, and topological evidence, and dynamic factors describing traffic interaction.

Scene context. Table 4 reveals substantially different performance patterns across PlanContext categories. For category A, contextual success decreases from 15.59% under low decision urgency to 3.22% under high urgency. B1, B2, L, and LC exhibit smaller variations across urgency levels. Geometric salience exhibits no consistent monotonic trend across categories. B1 reaches its highest performance at the medium level, C shows higher performance at greater salience, whereas LC

Table 4: Associations between scene context factors and PlanContext performance.

	Decision Urgency			Geometric Salience			Topological Evidence
Category	Low	Medium	High	Low	Medium	High	Δq [95% CI]
A	15.59	11.28	3.22	6.66	12.34	11.02	$+13.15^{*}$ [10.62, 15.72]
B1	64.22	61.90	62.63	55.74	77.43	55.49	$+35.67^{*}$ [21.22, 49.46]
B2	73.50	75.06	71.29	73.10	75.36	71.39	-13.45^{*} [-15.16, -11.78]
C	23.74	43.39	37.79	29.96	36.09	38.86	-20.77^{*} [-28.05, -13.39]
L	0.49	0.43	0.42	0.46	0.43	0.46	$+0.08$ [-0.07, 0.22]
LC	87.72	84.21	84.96	88.33	84.97	83.61	-2.51 [-7.62, 3.13]

- Decision Urgency and Geometric Salience report contextual success (%) under within-category Low/Medium/High stratification.
- Topological Evidence reports Strong-minus-Weak contextual success.
- * indicates that the 95% scene-token bootstrap confidence interval excludes zero.
- Detailed evidence definitions are provided in the Appendix.

shows the opposite tendency. Topological evidence provides a clearer distinction. Stronger target-branch directional evidence is associated with 13.15 and 35.67 percentage-point higher contextual success for A and B1, respectively, whereas B2 and C show 13.45 and 20.77 percentage-point lower success under their corresponding strong structural cues. These results show that road context provides different forms of decision evidence depending on the behavioral requirement.

Table 5: Associations between traffic interaction factors and PlanContext performance.

Category	Lead Vehicle	TTC Risk	Target Occupancy	Crossing Conflict	Traffic Density
A	-4.88	-9.08^{*}	-9.94^{*}	-0.18	$+5.31^{*}$
B1	$+17.47$	-5.15	-6.00	$+12.39$	-0.88
B2	$+9.05^{*}$	$+6.29^{*}$	$+5.54^{*}$	-4.45^{*}	$+12.10^{*}$
C	$+3.58$	-10.36	-10.02^{*}	-11.51^{*}	-9.75^{*}
L	-0.50^{*}	-0.26^{*}	-0.39^{*}	$+0.20$	-0.13^{*}
LC	-7.62^{*}	-8.62^{*}	-6.88^{*}	-9.56^{*}	-7.82^{*}

- Values report changes in contextual success (pp), computed as comparison minus reference.
- * indicates that the 95% scene-token bootstrap confidence interval excludes zero.

Traffic interaction. Table 5 shows similarly category-dependent associations between dynamic traffic conditions and contextual success. For Lead Vehicle and TTC, we compare $\leq$ 10 m with $>$ 25 m and $<$ 3 s with $\geq$ 6 s, respectively. Target Occupancy and Crossing Conflict compare $\geq$ 2 actors with 0 actors, while Traffic Density compares High with Low. A shows the largest performance differences across TTC risk and target-side occupancy, while C shows lower contextual success under greater target-side occupancy and crossing conflicts. LC exhibits consistent performance degradation across all five traffic factors, with lower success under more demanding dynamic conditions. For B2, closer lead vehicles, higher TTC risk, and greater target-side occupancy are associated with higher contextual success. Success increases by 12.10 percentage points from low- to high-density traffic, while crossing conflicts are associated with a decrease of 4.45 percentage points. These associations differ from those observed in LC, where contextual success decreases across all five traffic factors.

5 Conclusion

We present PlanContext, a benchmark for evaluating in-context driving decision-making through paired navigation interventions under fixed observations and vehicle history. Across 9,561 scene–condition instances and 23 end-to-end driving models, we find that strong conventional driving quality does not consistently correspond to strong in-context decision-making. Our analyses further show that model performance varies across different decision requirements and environmental contexts. These results indicate that in-context driving decision-making captures a capability dimension not fully reflected by conventional driving-quality metrics. By evaluating intention–context–decision consistency across route-level and lane-level scenarios, PlanContext provides a structured benchmark for assessing how end-to-end driving models respond to high-level intentions under different driving contexts.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and revision, translation, and consistency checks of reported numerical results. They also assisted with source searches, LaTeX editing, and table formatting. The authors reviewed the AI-assisted content, checked factual claims against the underlying experimental results and cited sources, and made the final decisions on the manuscript. The authors take responsibility for the accuracy and integrity of all content in this work.

ETHICS STATEMENT

PlanContext evaluates in-context driving decision-making using existing driving scenes from NAVSIM (Dauner et al., 2024) and publicly released or official model assets. All experiments are conducted offline, without interventions on real vehicles or road users. The human validation procedure involves expert inspection of existing scenes and model predictions, as described in Appendix D.9. Responsible reuse of the benchmark requires adherence to the source data licenses and protection of privacy in driving-scene imagery.

REPRODUCIBILITY STATEMENT

The paired navigation intervention protocol and contextual evaluation framework are specified in Section 3. Appendices A, B, and C detail scenario construction and filtering, category-specific evaluation rules, geometric thresholds, safety criteria, and metric aggregation. The experimental setup and model evaluation protocol are described in Section 4.1 and Appendix D.2. The scene-level bootstrap procedure and metric robustness analyses are documented in Appendices D.3 and D.7. These specifications support consistent reconstruction of the benchmark and its evaluation procedure. We plan to publicly release the PlanContext benchmark toolkit, including evaluation code and supporting documentation.

REFERENCES

Sining Ang, Yuguang Yang, Canyu Chen, and Yan Wang. Clover: Closed-loop value estimation and ranking for end-to-end autonomous driving planning. *arXiv preprint arXiv:2605.15120*, 2026.

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 *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Kunlin Cai, Rui Song, Jinghuai Zhang, Kaiyuan Zhang, Pranav Bodapati, Alicia Yu, Fnu Suya, Mohammad Rostami, Jiaqi Ma, and Yuan Tian. What-if world: A causal benchmark for general world models in embodied scenarios. *arXiv preprint arXiv:2605.27589*, 2026.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, et al. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218*, 2025.

Canyu Chen, Yuguang Yang, Zhewen Tan, Yizhi Wang, Ruiyi Zhan, Haiyan Liu, Xuanyao Mao, Jason Bao, Xinyue Tang, Linlin Yang, et al. Devil is in narrow policy: Unleashing exploration in driving vla models. *arXiv preprint arXiv:2603.06049*, 2026.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024.

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.

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 4693–4700. IEEE, 2018.

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.

Simon Gerstenecker, Andreas Geiger, and Katrin Renz. Fail2drive: Benchmarking closed-loop driving generalization. *arXiv preprint arXiv:2604.08535*, 2026.

Kaiser Hamid, Can Cui, and Nade Liang. Icr-drive: Instruction counterfactual robustness for end-to-end language-driven autonomous driving. *arXiv preprint arXiv:2604.05378*, 2026.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE, 2023.

Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends in Computer Graphics and Vision*, 12(1-3):1–308, 2020.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8306–8316. IEEE, 2023.

Ellington Kirby, Alexandre Boulch, Yihong Xu, Yuan Yin, Gilles Puy, Éloi Zablocki, Andrei Bursuc, Spyros Gidaris, Renaud Marlet, Florent Bartoccioni, et al. Driving on registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32058–32069, 2026.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27137–27146. IEEE, 2025.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 7890–7911, 2026.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12037–12047. IEEE, 2025.

Jinghui Lu, Jiayi Guan, Zhijian Huang, Jinlong Li, Guang Li, Lingdong Kong, Yingyan Li, Han Wang, Shaoqing Xu, Yuechen Luo, et al. Onevl: One-step latent reasoning and planning with vision-language explanation. *arXiv e-prints*, pp. arXiv–2604, 2026.

Ishaan Rawal, Shubh Gupta, Yihan Hu, and Wei Zhan. Nord: A data-efficient vision-language-action model that drives without reasoning. *arXiv preprint arXiv:2602.21172*, 2026.

Shuyao Shang, Bing Zhan, Yunfei Yan, Yuqi Wang, Yingyan Li, Yasong An, Xiaoman Wang, Jierui Liu, Lu Hou, Lue Fan, et al. Dynvla: Learning world dynamics for action reasoning in autonomous driving. *arXiv preprint arXiv:2603.11041*, 2026.

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. In *European Conference on Computer Vision*, pp. 446–463. Springer, 2026.

Linhan Wang, Zichong Yang, Chen Bai, Guoxiang Zhang, Xiaotong Liu, Xiaoyin Zheng, Xiao-Xiao Long, Chang-Tien Lu, and Cheng Lu. Drive-jepa: Video jepa meets multimodal trajectory distillation for end-to-end driving. *arXiv preprint arXiv:2601.22032*, 2026a.

Xiyang Wang, Xinlin Wang, Tingguang Zhou, Gong Chen, Xingtai Gui, Zhi Xu, Xiaolei Wu, Feiyang Tan, Hangning Zhou, and Mu Yang. Chainflow-vla: Causal flow planning with vision-language models. *arXiv preprint arXiv:2605.23270*, 2026b.

Hanfeng Wu, Marlon Steiner, Michael Schmidt, Alvaro Marcos-Ramiro, and Christoph Stiller. Closing the navigation compliance gap in end-to-end autonomous driving. *arXiv preprint arXiv:2512.10660*, 2025.

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1602–1611. IEEE, 2025.

Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 157518–157556, 2026.

Yifang Xu, Jiahao Cui, Zhihao Zhu, Hanlin Shang, Shan Luan, Mingwang Xu, Feipeng Cai, Neng Zhang, Yaoyi Li, Jia Cai, et al. Wam-flow: Parallel coarse-to-fine motion planning via discrete flow matching for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24918–24928, 2026.

Yibo Yuan, Jiacheng Fu, Jiangtong Zhu, Yi Li, Jianhua Han, Meng Tian, Zhuohan Liu, Zhiwei Xiong, Hang Xu, Jianwu Fang, et al. Suv: Future scene understanding as video generation for end-to-end driving. *arXiv preprint arXiv:2608.03084*, 2026.

Zongzheng Zhang, Jijun Wang, Saining Zhang, Shuo Wang, Yiru Wang, Hai Yang, Yang Chen, Yuwen Heng, Hao Sun, Anqing Jiang, et al. Plug-and-play traffic element awareness for end-to-end autonomous driving. *arXiv preprint arXiv:2608.18035*, 2026.

Zongchuang Zhao, Xin Zhou, Tianyang Xu, Zhengyang Sun, Kaixuan Zhou, Honglin Li, Dingkang Liang, and Xiang Bai. Simwam: A simple world action model for end-to-end autonomous driving. *arXiv preprint arXiv:2608.07468*, 2026.

Yinan Zheng, Tianyi Tan, Bin Huang, Enguang Liu, Ruiming Liang, Jianlin Zhang, Jianwei Cui, Guang Chen, Kun Ma, Hangjun Ye, et al. Unleashing the potential of diffusion models for end-to-end autonomous driving. *arXiv preprint arXiv:2602.22801*, 2026.

Xin Zhou, Zongchuang Zhao, Zhibo Yang, Mingsheng Li, Humen Zhong, Shuai Bai, Du Chu, Ruizhe Chen, Zhaohai Li, Jun Tang, et al. Qwen-drive-1.0: An initial step towards a vision-language foundation model for autonomous driving. *arXiv preprint arXiv:2609.00111*, 2026a.

Zewei Zhou, Tianhui Cai, Seth Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *Advances in Neural Information Processing Systems*, 38: 27920–27956, 2026b.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Diffusiondrivev2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.

Appendix

In the appendix, we provide additional details on the construction, evaluation, and analysis of PlanContext that are omitted from the main paper for brevity. The appendix is organized as follows:

- Sec. A provides detailed scenario mining, filtering, and category assignment criteria for the six PlanContext contexts.
- Sec. B presents the category-specific behavioral evaluation rules, including road-relative trajectory comparison, path-change detection, GT-relative suppression, and continuity evaluation for ongoing maneuvers.
- Sec. C describes road compliance, dynamic safety, and the construction of the final PlanContext metrics.
- Sec. D provides additional experimental details, statistical analyses, implementation information, and supplementary results.
- Sec. E presents qualitative examples across contextual categories and additional cases for ten evaluated models.

A Scenario Construction Details

This section provides the detailed scenario mining, filtering, and category assignment criteria used to construct PlanContext. The construction process first removes scenes with insufficient geometric or motion evidence. It then identifies reliable route branches and adjacent lane structures from the local map. Each retained scene–condition instance is assigned to one of the six contextual categories according to a fixed priority order. This procedure is designed to ensure that each category corresponds to a sufficiently clear contextual requirement.

A.1 Common Validity Filtering

Each candidate instance contains an original navigation intention and an alternative navigation intention selected from *left*, *straight*, and *right*. We retain only valid intention pairs for which the initial road, local map topology, logged trajectory, and the category-required reference geometry and permitted road regions can be reliably reconstructed. A target reference path is required only for categories where a valid target branch or adjacent lane is expected to exist.

We additionally remove near-stationary scenes using the future motion of the logged ego trajectory. Let

$$D_{\mathrm{GT}}^{4s} = \sum_{t} \left\| p_t^{\mathrm{GT}} - p_{t-1}^{\mathrm{GT}} \right\|_2 \tag{10}$$

denote the cumulative displacement of the logged trajectory over the future four-second horizon. A candidate is retained only when

$$D_{\mathrm{GT}}^{4s} \geq 2.0 \text{ m}. \tag{11}$$

This motion filtering removes 598 near-stationary candidates from the initial 12,840 candidates and leaves 12,242 candidates for subsequent contextual screening. Scenes in which the road structure, route branch, adjacent lane, or maneuver stage cannot be determined reliably are marked as ambiguous, transition, or pending and are excluded from the main benchmark.

A.2 Route Branch Construction and Reliability

For route-level categories, we search forward from the current road for the first decision point with multiple outgoing edges. The forward search range is limited to 200 m. Each candidate branch is then extended for up to 100 m after the decision point.

Candidate branches are classified according to their terminal heading relative to the reference heading at the decision point. Let $\Delta\theta$ denote this relative heading difference. A branch is classified as straight when

$$|\Delta\theta| \leq 20^{\circ}. \tag{12}$$

It is classified as left when

$$\Delta\theta \geq 30^\circ, \tag{13}$$

and as right when

$$\Delta\theta \leq -30^\circ. \tag{14}$$

Branches falling between these angular regions are considered ambiguous and are not used for reliable category assignment.

The original route branch is selected primarily according to route overlap. When multiple branches have identical route overlap, the logged trajectory is used only as a secondary geometric cue. The closest branch is selected only when its mean distance to the logged trajectory is at least 0.75 m smaller than that of the second closest branch. Otherwise, the instance is marked as having weak GT-based tie-breaking evidence and is excluded.

The original and alternative paths are sampled at intervals of 0.5 m. Reliable separation is searched only within their common valid path extent after the decision point, with the maximum search range defined as

$$s_{\max} = \min(L_N, L_F, 50\ \mathrm{m}), \tag{15}$$

where L_N and L_F denote the valid lengths of the original and alternative paths.

Within this range, the two paths are considered stably separated when three consecutive sampled positions satisfy at least one of the following conditions:

$$\left\|p^N(s) - p^F(s)\right\|_2 \geq 1.5\ \mathrm{m}, \tag{16}$$

or

$$|\Delta\theta(s)| \geq 12^\circ. \tag{17}$$

The earliest position at which this condition is continuously satisfied is recorded as the reliable separation position for the corresponding route pair.

A requested navigation intention may correspond to multiple alternative branches. In this case, the route evidence is considered reliable only when every alternative branch associated with the requested intention exhibits stable separation from the selected original branch. For categories A and B1, a valid arrival time must also be available for every alternative branch. If any branch lacks stable separation or a valid arrival estimate, the robust route evidence is treated as unavailable and the instance cannot be assigned to A or B1.

For an instance with multiple valid alternative branches, we use the latest reliable separation distance and the latest estimated arrival time:

$$d_{\mathrm{split}}^{\mathrm{robust}} = \max_j d_{\mathrm{split},j}, \tag{18}$$

$$t_{\mathrm{arrival}}^{\mathrm{robust}} = \max_j t_{\mathrm{arrival},j}. \tag{19}$$

Here, j indexes all alternative branches associated with the requested navigation intention. The arrival time of each branch is estimated from the cumulative future distance of the logged trajectory. Linear interpolation is applied between neighboring trajectory samples, which are spaced at 0.5 s intervals.

A.3 Category Assignment Priority

Some scenes satisfy more than one preliminary geometric condition. We therefore apply a fixed category assignment order to avoid overlapping definitions.

A scene is first assigned to category C when strict post-commitment evidence is detected. A scene that starts on a lane connector is excluded when strict commitment evidence is unavailable. If a reliable route divergence can be reached within 2.0 s, the instance is assigned to category A. Cases with an arrival time between 2.0 s and 4.5 s are treated as transition cases and are excluded.

If an active lateral maneuver is already present, the instance is assigned to category LC. Among the remaining cases, a reliable route divergence with an arrival time between 4.5 s and 8.0 s forms a category B1 candidate. A scene is assigned to category L when the ego vehicle remains stably within its current lane and a geometrically feasible adjacent target lane is available. Category B2 is used

when no reliable nearby route branch and no feasible adjacent target lane support the alternative intention. All remaining cases are marked as ambiguous or pending and are excluded from the main benchmark.

A.4 Category-Specific Construction Criteria

Category A: Actionable Route Change. Category A contains scenes in which the alternative route has entered the current decision window and should produce a target-directed response within the four-second prediction horizon. A reliable alternative route branch must satisfy

$$0 < t_{\text{arrival}}^{\text{robust}} \leq 2.0 \text{ s}. \tag{20}$$

The ego vehicle must not have strictly committed to the original route. Cases beginning on a lane connector are also excluded when sufficient commitment evidence is unavailable. The interval

$$2.0 < t_{\text{arrival}}^{\text{robust}} < 4.5 \text{ s} \tag{21}$$

is treated as a transition region and does not contribute to either A or B1.

Category B1: Delayed Route Change. Category B1 represents scenes in which the alternative route is feasible in the near future but should not yet induce a directional commitment within the current prediction horizon. The reliable arrival time must satisfy

$$4.5 \leq t_{\text{arrival}}^{\text{robust}} \leq 8.0 \text{ s}. \tag{22}$$

We additionally require sufficient remaining distance between the reliable branch separation and the future motion covered by the logged trajectory:

$$d_{\text{split}}^{\text{robust}} - D_{\text{GT}}^{4s} \geq 8.0 \text{ m}. \tag{23}$$

The scene must not contain a feasible adjacent lane that could already provide a valid immediate interpretation of the alternative navigation intention.

Category B2: Locally Invalid Target. Category B2 contains alternative left or right intentions that do not correspond to a locally reachable target in the requested direction. The scene must contain neither a reliable nearby route branch in the requested direction nor a geometrically feasible adjacent target lane within the same roadblock.

We conservatively exclude candidates for which the local map still provides evidence of a potentially reachable target. In particular, a candidate is excluded from B2 whenever at least one branch associated with the alternative direction is detected and its decision point lies within the four-second logged travel distance,

$$N_{\text{replacement}} > 0 \tag{24}$$

and

$$d_{\text{decision}} \leq D_{\text{GT}}^{4s}. \tag{25}$$

This exclusion is applied even when the complete route-pair reliability test is not satisfied. Such cases are excluded from B2 and retained as boundary cases for analysis.

Candidates are also excluded when an adjacent lane exists but fails the required geometric continuity checks. These exclusions reduce ambiguity in the B2 set by removing cases with residual evidence of a potentially reachable target.

Category C: Post-Commitment Route Change. Category C represents scenes in which the alternative route was available at an earlier decision point and the ego vehicle has already passed the relevant divergence and committed to the original route.

Candidate detection searches backward for up to 80 m to locate the nearest previous decision point with multiple outgoing branches. The current branch must agree with the original left or right navigation intention. The alternative direction must also have existed at that historical decision point. The original and alternative branches must exhibit reliable geometric separation. In addition,

the ego heading change over approximately the previous two seconds must be consistent with the original turning direction and must reach at least 10°.

We measure the post-commitment progress along the selected committed branch as

$$d_{\text{commit}} = s_{\text{ego}}^{\text{committed}} - s_{\text{divergence}}, \tag{26}$$

where $s_{\text{ego}}^{\text{committed}}$ denotes the ego progress measured from the beginning of the selected committed branch and $s_{\text{divergence}}$ denotes the stable separation position measured along the same branch.

Candidate detection requires

$$d_{\text{commit}} \geq 3.0 \text{ m}. \tag{27}$$

The main benchmark uses the stricter criterion

$$d_{\text{commit}} \geq 5.0 \text{ m}. \tag{28}$$

Cases with

$$3.0 \text{ m} \leq d_{\text{commit}} < 5.0 \text{ m} \tag{29}$$

are excluded from the main benchmark and retained as boundary cases for analysis.

Category L: Feasible Adjacent Lane Change. Category L contains scenes in which the ego vehicle has not started a lateral maneuver and the adjacent target lane in the requested direction is currently reachable. The ego vehicle must be within 1.0 m of the centerline of the current lane. The nearest lane associated with each of the most recent three history samples must also remain the current lane.

For the target adjacent lane, the relative heading, lateral distance, and parallel overlap must satisfy

$$|\Delta\theta_{\text{lane}}| \leq 15^\circ, \tag{30}$$

$$d_{\text{lateral}} \leq 8.0 \text{ m}, \tag{31}$$

and

$$L_{\text{parallel}} \geq 20.0 \text{ m}. \tag{32}$$

The available map interface does not provide complete legal attributes for every lane boundary. We therefore interpret these cases as topologically feasible adjacent lane changes whose boundary legality cannot always be independently verified. Cases with discontinuous geometry or unstable current lane assignment are excluded.

Category LC: Ongoing Lateral Maneuver. Category LC contains scenes in which a persistent lateral motion is already observable before the navigation intention is changed. We use the ego history from $t = -2.0$ s to $t = 0$ s with a sampling interval of 0.5 s.

An active lateral direction is detected through one of two motion patterns. In the first pattern, the historical trajectory starts at least 0.5 m closer to an adjacent lane than to the current lane,

$$d_{\text{adj}}(-2s) + 0.5 < d_{\text{current}}(-2s). \tag{33}$$

The trajectory must then move at least 0.75 m farther from that adjacent lane,

$$d_{\text{adj}}(0) - d_{\text{adj}}(-2s) \geq 0.75 \text{ m}, \tag{34}$$

while moving at least 0.75 m closer to the current lane,

$$d_{\text{current}}(-2s) - d_{\text{current}}(0) \geq 0.75 \text{ m}. \tag{35}$$

In the second pattern, the trajectory moves from the current lane toward an adjacent lane. It must become at least 0.75 m closer to the adjacent lane while moving at least 0.75 m farther from the current lane.

The relation between the alternative navigation intention and the detected active direction is recorded as *continue*, *reverse*, or *cancel*. These relations are used for supplementary analysis and for the continuity evaluation described in Sec. B. They are not treated as separate benchmark categories.

Table 6: Scenario filtering statistics for PlanContext.

Stage	Instances	Description
Initial candidates	12,840	Original scene–condition candidates
After motion filtering	12,242	Remove near-stationary instances
Final benchmark	9,561	Apply final context-aware filtering

A.5 Scenario Filtering Statistics

Table 6 summarizes the main filtering stages used to construct PlanContext. Starting from 12,840 scene–condition candidates, motion filtering retains 12,242 instances. The subsequent context-aware screening and category-specific validity checks yield the final 9,561 instances used for evaluation.

The final benchmark contains 334 A, 251 B1, 5,285 B2, 580 C, 2,716 L, and 395 LC instances.

B Category-Specific Evaluation Rules

This section provides the detailed trajectory comparison and category-specific behavioral evaluation rules used by PlanContext. Given the original prediction T_i^N and the prediction under the alternative navigation intention T_i^F, we first represent both trajectories with respect to the reference geometry required by each evaluation component. We then determine their topological outcomes, identify persistent behavioral changes induced by the navigation intervention, and evaluate whether these changes satisfy the contextual requirement of each category.

B.1 Notation and Reference Geometries

Each prediction contains eight future ego states sampled at 0.5 s intervals over a four-second horizon. Predicted poses are interpreted at the rear axle center and transformed into the global SE(2) coordinate frame.

We use two reference geometries for different evaluation purposes. The original-route reference is used to characterize the path-level difference between the original and alternative predictions. The logged GT trajectory provides a local reference for the suppression tests of B1, B2, and C and for the continuity evaluation of LC. The GT reference represents the natural road evolution observed in the scene and is not treated as a trajectory target to be imitated.

For the common path-change analysis, each prediction is projected onto the original-route reference. For a projected state, we record the longitudinal road progress s, signed lateral offset d, and relative heading

$$\theta_r = \mathrm{wrap}\left(\theta_{\mathrm{ego}} - \theta_{\mathrm{road}}\right). \tag{36}$$

The original and alternative predictions are aligned at common road-progress values. We define

$$\Delta d(s) = d^F(s) - d^N(s) \tag{37}$$

and

$$\Delta\theta(s) = \mathrm{wrap}\left(\theta_r^F(s) - \theta_r^N(s)\right). \tag{38}$$

This progress-based alignment reduces the influence of longitudinal speed differences when comparing behavioral changes induced by the navigation intervention.

The original-route and target-route references are also used to characterize the terminal topology of each prediction. We denote them as $\mathcal{R}_o$ and $\mathcal{R}_t$. The GT reference used for suppression and continuity evaluation is denoted as $\mathcal{R}_{GT}$.

B.2 Evaluation Structure

The category-specific criteria are constructed from a shared set of trajectory-level primitives, including route-topology assignment, road-relative trajectory change, GT-relative residual change,

full-vehicle road compliance, and dynamic safety. Each contextual category selects the evidence required by its behavioral objective.

Categories A and L evaluate whether a feasible alternative intention is executed. Categories B1, B2, and C evaluate whether the alternative navigation condition induces an unnecessary persistent deviation from the locally appropriate behavior. Category LC evaluates the continuity of an ongoing lateral maneuver.

This factorized design keeps the underlying geometric measurements shared across categories while allowing the decision criterion to reflect the corresponding contextual requirement. Continuous trajectory evidence is evaluated using fixed decision thresholds and boundary regions. Clear evidence contributes to the category-specific decision outcome, while boundary-level evidence is retained for uncertainty-aware evaluation. All category definitions and thresholds are fixed and shared across models.

B.3 Topological Outcome Assignment

We determine the topological outcome of each predicted trajectory from its terminal portion. Let d_o and d_t denote the distances from the trajectory tail to the original-route reference $\mathcal{R}_o$ and target-route reference $\mathcal{R}_t$.

When both reference distances are available, a trajectory is assigned to the target topology when

$$d_t + 0.5 < d_o. \tag{39}$$

When

$$d_o + 0.5 < d_t, \tag{40}$$

the trajectory is assigned to the original topology for categories A and L. The same geometric condition is labeled as committed for category C and as active maneuver for category LC.

When the difference between d_o and d_t is smaller than 0.5 m, category B1 uses the shared label. The corresponding outcome is marked as ambiguous for the remaining categories.

Reference geometry can be incomplete for some local configurations. When only d_o is available, the prediction is assigned to the original topology for A, B1, B2, and L. The corresponding labels are committed for C and active maneuver for LC.

When d_o is unavailable, category B2 uses an auxiliary state denoted as *other-local*. The remaining categories use the ambiguous state. The *other-local* state indicates insufficient comparative evidence to assign the prediction to either the original topology or a supported target topology. It does not indicate road compliance or traffic-rule legality. Road compliance is evaluated independently using the full vehicle footprint in Sec. C.

The resulting topology variable is denoted by τ_i^N or τ_i^F for the original and alternative predictions.

B.4 Persistent Path-Change Detection

We next determine whether changing the navigation intention produces a persistent change in the predicted path. Let the trajectory tail contain the final three valid aligned samples. We compute

$$\widetilde{d}_{\text{tail}} = \text{median}\left(\Delta d_{\text{tail}}\right), \tag{41}$$

$$d_{\text{end}} = \Delta d(s_{\text{end}}), \tag{42}$$

$$d_{75} = Q_{0.75}\left(|\Delta d|\right), \tag{43}$$

and

$$\widetilde{\theta}_{\text{tail}} = \text{median}\left(\Delta \theta_{\text{tail}}\right). \tag{44}$$

For samples satisfying $|\Delta d| \geq 0.15$ m, we compute the sign-consistency ratio r_{sign}. This quantity measures the fraction of valid samples whose lateral-change sign agrees with the dominant direction of the trajectory tail.

A strong path change is identified when at least one of the following conditions is satisfied.

1. The original and alternative predictions are assigned to different road topologies.
2. At least four aligned samples are available and

$$|\widetilde{d}_{\text{tail}}| \geq 0.85 \text{ m}, \tag{45}$$

$$|d_{\text{end}}| \geq 0.85 \text{ m}, \tag{46}$$

$$d_{75} \geq 0.60 \text{ m}, \tag{47}$$

and

$$r_{\text{sign}} \geq 0.80. \tag{48}$$

3. The tail heading and lateral differences satisfy

$$|\widetilde{\theta}_{\text{tail}}| \geq 8.5^\circ, \tag{49}$$

$$|\widetilde{d}_{\text{tail}}| \geq 0.60 \text{ m}, \tag{50}$$

and

$$r_{\text{sign}} \geq 0.80. \tag{51}$$

We additionally define a boundary-evidence region for diagnostic review. The lateral criterion uses 0.65 m for the tail and endpoint differences, 0.40 m for d_{75}, and 0.70 for the sign-consistency ratio. The heading-based criterion uses a heading difference of 6.5°, a lateral difference of at least 0.40 m, and a sign-consistency ratio of 0.70.

If no strong path change is detected and

$$\left|s^F_{\text{end}} - s^N_{\text{end}}\right| \geq 1.0 \text{ m}, \tag{52}$$

the response is labeled as a progress-only change. Responses without a strong path change or sufficient longitudinal difference are labeled as no change. Cases with insufficient common progress or unavailable reference geometry are labeled as geometry ambiguous.

B.5 Direction of the Path Change

For a persistent path change, we next determine whether the change is directed toward the alternative target, toward the original route, or toward an unrelated direction.

We first apply topology-based decisions. A transition from a non-target topology under the original condition to the target topology under the alternative condition is labeled as a topology switch. If the original prediction is assigned to the target topology and the alternative prediction leaves the target topology, the change is labeled as original-directed.

For the remaining persistent path changes, let

$$g(T) = d(T, \mathcal{R}_o) - d(T, \mathcal{R}_t) \tag{53}$$

denote the relative target preference of trajectory T. We define

$$\Delta g = g(T^F) - g(T^N). \tag{54}$$

When the required route distances are available, the change is labeled as target-directed when

$$\Delta g \geq 1.2 \text{ m}. \tag{55}$$

The change is labeled as original-directed when

$$\Delta g \leq -1.2 \text{ m}. \tag{56}$$

All remaining persistent directional changes are labeled as unrelated-directional.

B.6 GT-Relative Suppression for B1, B2, and C

Categories B1, B2, and C require the alternative navigation intention to avoid inducing an unnecessary persistent path deviation. The logged trajectory is used as a reference for the natural local road evolution. It is not treated as the unique correct trajectory under the alternative navigation condition.

Both predictions are projected onto the GT reference $\mathcal{R}_{GT}$. We define the tail lateral residuals as

$$r_d^N = \left|\text{median}\left(d_{\text{tail}}^N\right)\right| \tag{57}$$

and

$$r_d^F = \left|\text{median}\left(d_{\text{tail}}^F\right)\right|. \tag{58}$$

The corresponding heading residuals are

$$r_\theta^N = \left|\text{median}\left(\theta_{\text{tail}}^N\right)\right| \tag{59}$$

and

$$r_\theta^F = \left|\text{median}\left(\theta_{\text{tail}}^F\right)\right|. \tag{60}$$

The additional deviation induced by the alternative navigation condition is

$$\Delta r_d = r_d^F - r_d^N \tag{61}$$

and

$$\Delta r_\theta = r_\theta^F - r_\theta^N. \tag{62}$$

An unnecessary persistent response is identified only when a strong path change has already been detected. Strong lateral evidence is present when

$$r_d^F \geq 1.5\ \text{m} \tag{63}$$

and

$$\Delta r_d \geq 1.0\ \text{m}. \tag{64}$$

Strong heading evidence is present when

$$r_\theta^F \geq 12^\circ \tag{65}$$

and

$$\Delta r_\theta \geq 10^\circ. \tag{66}$$

Two additional conditions provide strong evidence of an unnecessary response. The first occurs when the original and alternative predictions lie on opposite sides of the GT reference and satisfy

$$r_d^F \geq 0.85\ \text{m} \tag{67}$$

and

$$r_d^N \geq 0.40\ \text{m}. \tag{68}$$

The second occurs when the alternative prediction switches from a non-target topology to the target topology.

We denote the resulting unnecessary-response indicator as U_i. The GT-relative evaluation is considered available when the required GT reference geometry and residual quantities can be computed. We denote this availability by G_i.

Boundary evidence is retained for diagnostic review. The GT-relative boundary thresholds use a lateral residual of at least 1.0 m with an additional increase of at least 0.5 m, or a heading residual of at least 8° with an additional increase of at least 6°. A sample is also retained for diagnostic review when the common path-change detector in Sec. B.4 produces only boundary evidence. Diagnostic boundary evidence does not automatically make the contextual decision incorrect.

B.7 Category-Specific Behavioral Correctness

For categories A, B1, B2, C, and L, contextual correctness combines the semantic validity of the two predictions with their paired behavioral relationship.

Category A. The original prediction is semantically valid when

$$S_i^N = [\tau_i^N = \text{original}], \tag{69}$$

and the alternative prediction is semantically valid when

$$S_i^F = [\tau_i^F = \text{target}]. \tag{70}$$

The paired behavioral relation is correct when the detected path change is target-directed or produces a topology switch:

$$P_i = [\text{target-directed or topology switch}]. \tag{71}$$

The contextual decision correctness is

$$C_i = S_i^N \wedge S_i^F \wedge P_i. \tag{72}$$

A missing response, an incorrect branch transition, or a persistent change in an unrelated direction is counted as incorrect.

Category B1. The original and alternative predictions are semantically valid when they remain on the shared approach or original route:

$$S_i^N = [\tau_i^N \in \{\text{shared}, \text{original}\}], \tag{73}$$

$$S_i^F = [\tau_i^F \in \{\text{shared}, \text{original}\}]. \tag{74}$$

The paired behavioral requirement is

$$P_i = G_i \wedge \neg U_i, \tag{75}$$

and the contextual decision correctness is

$$C_i = S_i^N \wedge S_i^F \wedge P_i. \tag{76}$$

Longitudinal slowing, braking, progress-only changes, and natural road curvature remain valid. Any unnecessary persistent path deviation induced by the alternative intention is counted as incorrect. In B1, the most direct failure mode is premature commitment toward a target branch that has not yet entered the current decision window.

Category B2. The original and alternative predictions are semantically valid when they do not enter the unsupported target topology:

$$S_i^N = [\tau_i^N \neq \text{target}], \tag{77}$$

$$S_i^F = [\tau_i^F \neq \text{target}]. \tag{78}$$

The paired behavioral requirement is

$$P_i = G_i \wedge \neg U_i, \tag{79}$$

which gives

$$C_i = S_i^N \wedge S_i^F \wedge P_i. \tag{80}$$

A lack of directional response is compatible with B2 but does not independently guarantee final success. Any unnecessary persistent path deviation induced by the alternative intention is counted as incorrect. This includes a persistent response toward an unsupported target and a strong unrelated directional deviation. Road compliance is evaluated separately in Sec. C.

Category C. The original and alternative predictions must remain consistent with the already committed route:

$$S_i^N = [\tau_i^N \in \{\text{committed}, \text{original}\}], \tag{81}$$

$$S_i^F = [\tau_i^F \in \{\text{committed}, \text{original}\}]. \tag{82}$$

The paired behavioral requirement is

$$P_i = G_i \wedge \neg U_i, \tag{83}$$

and

$$C_i = S_i^N \wedge S_i^F \wedge P_i. \tag{84}$$

The model may slow down, stop, or modify its longitudinal progress. Any unnecessary persistent deviation induced by the alternative intention is counted as incorrect. Typical failures include reversal toward a previously missed alternative branch, an additional pull toward that branch, and a strong unrelated directional deviation.

Category L. The original prediction is required to remain in the original lane:

$$S_i^N = [\tau_i^N = \text{original}]. \tag{85}$$

The alternative prediction is required to enter the target adjacent lane:

$$S_i^F = [\tau_i^F = \text{target}]. \tag{86}$$

The paired behavioral relation is correct when the path change is target-directed or produces a topology switch:

$$P_i = [\text{target-directed or topology switch}]. \tag{87}$$

The contextual decision correctness is

$$C_i = S_i^N \wedge S_i^F \wedge P_i. \tag{88}$$

A small lateral or heading response that remains within the current lane is not counted as a completed lane change.

B.8 LC Continuity Evaluation

Category LC evaluates whether changing the navigation intention disrupts an ongoing lateral maneuver. It does not use independent semantic labels for the original and alternative predictions. The evaluation produces one of three states:

$$P_i^{LC} \in \{\text{correct}, \text{incorrect}, \text{uncertain}\}. \tag{89}$$

Both predictions are projected onto the GT reference $\mathcal{R}_{GT}$. At each future time step t, we define the signed lateral residuals

$$d_t^N = d(T_t^N, \mathcal{R}_{GT}), \tag{90}$$

$$d_t^F = d(T_t^F, \mathcal{R}_{GT}), \tag{91}$$

and the corresponding heading residuals

$$\theta_t^N = \text{wrap}\left(\theta(T_t^N) - \theta_{GT}(s_t)\right), \tag{92}$$

$$\theta_t^F = \text{wrap}\left(\theta(T_t^F) - \theta_{GT}(s_t)\right). \tag{93}$$

The absolute residual measures deviation from the natural GT road evolution. The difference between the alternative and original residuals is used to distinguish intervention-induced changes from deviations that are already present in the original prediction.

The continuity protection window contains the first three predicted states at 0.5 s, 1.0 s, and 1.5 s.

Early Additional Deviation. A strong early lateral deviation is detected when at least two states within the protection window satisfy

$$|d_t^F| \geq 1.0 \text{ m} \tag{94}$$

and

$$|d_t^F| - |d_t^N| \geq 0.75 \text{ m}. \tag{95}$$

A strong early heading deviation is detected when at least two states satisfy

$$|\theta_t^F| \geq 10^\circ \tag{96}$$

and

$$|\theta_t^F| - |\theta_t^N| \geq 8^\circ. \tag{97}$$

The corresponding boundary thresholds are 0.75 m and 0.50 m for lateral deviation and 8° and 6° for heading deviation.

Persistent Reversal Within the Protection Window. Using the active maneuver direction detected from the history, we define the opposite lateral direction as the reversal direction. Let d_{reverse}^N and d_{reverse}^F denote the projected displacement components along this reversal direction.

A strong persistent reversal is detected when at least two states within the protection window satisfy

$$d_{\text{reverse}}^F \geq 1.0 \text{ m} \tag{98}$$

and

$$d_{\text{reverse}}^F - d_{\text{reverse}}^N \geq 0.75 \text{ m}. \tag{99}$$

The corresponding boundary thresholds are 0.75 m and 0.50 m.

Reversal Beginning Near the End of the Protection Window. For the *reverse* relation, we additionally detect sustained reversals that begin near the end of the protection window.

Let $\Delta d_{\text{rev}}(t)$ denote the additional lateral deviation of the alternative prediction relative to the original prediction along the reversal direction at future time t. Similarly, $\Delta\theta_{\text{rev}}(t)$ denotes the additional wrapped heading deviation along the reversal direction. Positive values indicate increasing deviation in the reversal direction.

The reversal consistency ratio r_{rev} is defined as the fraction of valid aligned future states whose signed heading change is consistent with the reversal direction.

Strong evidence requires

$$\Delta\theta_{\text{rev}}(1.5s) \geq 12^\circ, \tag{100}$$

$$\Delta\theta_{\text{rev}}(2.0s) \geq 18^\circ, \tag{101}$$

$$\Delta\theta_{\text{rev}}(2.0s) - \Delta\theta_{\text{rev}}(1.5s) \geq 3^\circ, \tag{102}$$

$$\Delta d_{\text{rev}}(2.0s) \geq 0.75 \text{ m}, \tag{103}$$

$$\text{median}_{t>2s}\, \Delta\theta_{\text{rev}}(t) \geq 20^\circ, \tag{104}$$

and

$$r_{\text{rev}} \geq 0.75. \tag{105}$$

The corresponding boundary thresholds are 8°, 12°, 2°, 0.50 m, 15°, and 0.75.

Unrelated Persistent Deviation. An unrelated-directional path change is treated as strong continuity-disruption evidence when the alternative prediction also exhibits a sufficiently large residual relative to the GT reference.

The lateral condition is

$$r_d^F \geq 1.5 \text{ m} \tag{106}$$

and

$$\Delta r_d \geq 1.0 \text{ m}. \tag{107}$$

The heading condition is

$$r_\theta^F \geq 12^\circ \tag{108}$$

and

$$\Delta r_\theta \geq 10^\circ. \tag{109}$$

An unrelated-directional path change satisfying either condition is counted as strong disruption evidence.

The corresponding boundary thresholds use a lateral residual of at least 1.0 m with an additional increase of at least 0.5 m, or a heading residual of at least 8° with an additional increase of at least 6°.

Final LC Status. The LC continuity status is assigned as

$$P_i^{LC} = \begin{cases} \text{incorrect}, & \text{if strong disruption evidence is present,} \\ \text{uncertain}, & \text{if only boundary evidence is present,} \\ \text{correct}, & \text{otherwise.} \end{cases} \tag{110}$$

A smooth response to the new navigation intention after the protection window is allowed. Large absolute deviation from the logged trajectory is retained as diagnostic information and does not independently determine the LC status.

For LC, contextual correctness is

$$C_i = 1 \tag{111}$$

when $P_i^{LC} = \text{correct}$ and

$$C_i = 0 \tag{112}$$

when $P_i^{LC} = \text{incorrect}$. Instances with $P_i^{LC} = \text{uncertain}$ are excluded from subsequent success-rate aggregation.

C Road Compliance, Safety, and Metric Details

This section describes the road-compliance and dynamic-safety evaluation used by PlanContext, followed by the aggregation of instance-level results into the final benchmark metrics. These components are evaluated separately from the contextual decision correctness defined in Sec. B.

C.1 Full-Vehicle Road Compliance

Road compliance is evaluated using the complete ego-vehicle footprint at every predicted future state. Each predicted pose is interpreted at the rear axle center. We use the dimensions of the Pacifica vehicle model, with a length of 5.176 m and a width of 2.297 m. The rear axle is located 1.461 m behind the geometric center of the vehicle footprint.

Let $\mathcal{B}(T_{i,t}^F)$ denote the rectangular vehicle footprint corresponding to the alternative prediction at future state t. Let Ω_i^F denote the road region permitted by the contextual requirement of instance i. A 0.35 m geometric buffer is applied to the permitted road region.

The road-compliance indicator is defined as

$$R_i^F = \prod_{t=1}^{8} \mathbb{1}\left[\mathcal{B}(T_{i,t}^F) \subseteq \Omega_i^F\right]. \tag{113}$$

A prediction passes the road-compliance evaluation only when the complete vehicle footprint is contained in the permitted region at all eight future states.

The permitted road region depends on the contextual category. Table 7 summarizes the corresponding regions under the original and alternative navigation conditions.

For A, the permitted region follows the original route under the original navigation condition and the target route under the alternative condition. For B1, both conditions remain within the original

Table 7: Permitted road regions for the six PlanContext categories.

Category	Original condition	Alternative condition
A	Original branch	Target branch
B1	Original approach branch	Original approach branch
B2	Local valid roads and adjacent roads	Local valid roads and adjacent roads
C	Committed original branch	Committed original branch
L	Current lane	Current and target adjacent lanes
LC	Maneuver-related local roads	Maneuver-related local roads

approach region because the alternative route has not yet entered the current decision window. B2 uses the locally available road region because no supported target route is defined. Category C follows the already committed original route. Category L uses the current lane under the original condition and the union of the current and target adjacent lanes under the alternative condition. Category LC uses the local road region associated with the ongoing lateral maneuver.

Each permitted corridor is extended sufficiently far to cover the longest predicted trajectory arc length, the vehicle front overhang, and an additional 5 m margin. For category A, a target branch that terminates within the local map is allowed to end at its valid road termination.

C.2 Dynamic and Rule-Based Safety

Dynamic safety is evaluated using the NAVSIM safety criteria on the alternative prediction. The evaluation includes collision avoidance, drivable-area compliance, driving-direction compliance, traffic-light compliance, time-to-collision, and history-based comfort.

Let $e^F_{i,m}$ denote the binary result of safety component m for instance i. The dynamic-safety indicator is

$$E^F_i = \bigwedge_m e^F_{i,m}. \tag{114}$$

The evaluated components are

$$m \in \{\text{collision}, \text{drivable}, \text{direction}, \text{traffic light}, \text{TTC}, \text{comfort}\}. \tag{115}$$

The overall safety condition combines the full-vehicle road-compliance check with the NAVSIM safety criteria:

$$\text{OverallSafety}_i = R^F_i \wedge E^F_i. \tag{116}$$

Road compliance and NAVSIM dynamic safety are kept as separate components because they characterize different failure modes. The former evaluates whether the complete predicted vehicle footprint remains within the context-specific permitted region. The latter evaluates dynamic interactions and rule-related safety conditions.

C.3 Instance-Level Contextual Success

Let C_i denote the contextual decision correctness defined in Sec. B. The final contextual success of instance i is

$$q_i = C_i \wedge R^F_i \wedge E^F_i. \tag{117}$$

A successful instance therefore requires a context-consistent decision response, road-compliant motion, and dynamic safety.

For additional analysis, we additionally define the correctness of the alternative-condition prediction as

$$A^F_i = \begin{cases} S^F_i, & k_i \in \{A, L\}, \\ S^F_i \wedge G_i \wedge \neg U_i, & k_i \in \{B1, B2, C\}, \\ \nVdash \left[P^{LC}_i = \text{correct}\right], & k_i = LC. \end{cases} \tag{118}$$

The corresponding alternative-condition success indicator is

$$q_i^{\mathrm{alt}} = A_i^F \wedge R_i^F \wedge E_i^F. \tag{119}$$

This metric evaluates whether the prediction under the alternative navigation condition is contextually valid, road compliant, and dynamically safe. For B1, B2, and C, alternative-condition validity additionally requires that the GT-relative suppression test is available and that no unnecessary persistent response is detected. For LC, the alternative-condition correctness is determined by the continuity status defined in Sec. B.8.

LC instances with $P_i^{LC} = \mathrm{uncertain}$ are excluded from the valid denominator of the alternative-condition success rate.

C.4 Class-Balanced Metric Aggregation

The six PlanContext categories contain substantially different numbers of instances. We therefore compute the contextual success rate independently within each category before aggregating across categories.

For category k, the contextual success rate is

$$S_k = \frac{1}{|\mathcal{D}_k|} \sum_{i \in \mathcal{D}_k} q_i, \tag{120}$$

where $\mathcal{D}_k$ denotes the valid scene–condition instances used for metric aggregation in category k. For LC, instances labeled as *uncertain* are excluded from $\mathcal{D}_{LC}$.

We aggregate categories A, B1, B2, and C for route-level evaluation and report categories L and LC separately for lane-level evaluation.

The Navigation Action Score (NAS) is the contextual success rate in category A, where the alternative route is immediately actionable. The Context Preservation Score (CPS) averages the contextual success rates across B1, B2, and C, where the alternative navigation intention should be deferred or rejected:

$$\mathrm{CPS} = \frac{S_{B1} + S_{B2} + S_C}{3}. \tag{121}$$

The Navigation Context Score (NCS) summarizes route-level in-context driving decision-making through the geometric mean of NAS and CPS:

$$\mathrm{NCS} = \sqrt{\mathrm{NAS} \cdot \mathrm{CPS}}. \tag{122}$$

CPS is a macro-average of the success rates in B1, B2, and C, with each category-level score assigned a weight of one third. The geometric mean penalizes imbalanced performance across NAS and CPS.

For lane-level in-context driving decision-making, the Lane Adaptation Score (LAS) and Maneuver Continuity Score (MCS) are the contextual success rates in L and LC, respectively. These scores are reported separately and are not included in NCS.

For comparison with conventional driving quality, matched PDMS is computed on the same 5,911 physical scenes used by PlanContext under the original navigation condition. Each physical scene contributes once to matched PDMS even when it is associated with multiple alternative navigation conditions.

D Additional Experimental Details and Analyses

This section provides additional details on the benchmark composition, model evaluation protocol, statistical analysis, contextual factor definitions, and robustness analyses used in PlanContext. The definitions of scenario construction, contextual correctness, road compliance, and metric aggregation are provided in Secs. A, B, and C.

Table 8: Composition of the PlanContext benchmark.

Category	Requirement	Instances
A	Follow	334
B1	Wait	251
B2	Reject	5,285
C	Reject	580
L	Follow	2,716
LC	Preserve	395
Total		9,561

D.1 Benchmark Composition

PlanContext is constructed from the NAVSIM navtest split. The final benchmark contains 9,561 scene–condition instances derived from 5,911 unique physical scenes. A physical scene may contribute multiple scene–condition instances when multiple alternative navigation intentions satisfy the construction criteria.

The six contextual categories contain 334 A, 251 B1, 5,285 B2, 580 C, 2,716 L, and 395 LC instances. Table 8 summarizes the benchmark composition.

The benchmark is category imbalanced because the six contextual conditions occur at different frequencies in NAVSIM. We therefore use the class-balanced aggregation defined in Sec. C.4 instead of a direct micro-average over all instances.

D.2 Model Evaluation Protocol

We evaluate 23 publicly available end-to-end driving models covering three representative paradigms. Vision and state to action models directly predict future trajectories from sensory observations or state inputs. Vision language action models incorporate vision language representations into driving policies. World model based action models explicitly predict or encode future scene dynamics before producing driving actions.

All models are evaluated using publicly released or official assets. Navigation conditions are explicitly involved during inference. Model outputs are standardized to eight future ego poses sampled over a four-second prediction horizon.

For each scene–condition instance, the model is evaluated twice. The first inference uses the original navigation intention c_i^N. The second inference uses the alternative navigation intention c_i^F. Sensor observations and vehicle history remain unchanged between the two evaluations. This paired protocol isolates the behavioral response to the navigation intervention while preserving the observed scene.

For conventional driving evaluation, we report PDMS on the full NAVSIM navtest split. We additionally compute matched PDMS, denoted as mPDMS, on the same 5,911 physical scenes used by PlanContext. mPDMS is evaluated under the original navigation condition. Each physical scene contributes once to mPDMS even when it produces multiple scene–condition instances in PlanContext.

The PlanContext evaluation covers all 9,561 scene–condition instances across the six categories. For each model, we report NAS, CPS, and NCS for route-level evaluation, LAS and MCS for lane-level evaluation, and the six category-wise contextual success rates.

D.3 Statistical Analysis

We estimate uncertainty using scene-level bootstrap resampling. The resampling unit is the unique physical scene token instead of an individual scene–condition instance. This design preserves the dependence among multiple navigation conditions constructed from the same physical scene.

Table 9: NCS and 95% scene-level bootstrap confidence intervals for the 23 evaluated models.

Model	NCS	95% CI
OneVL	39.30	[34.73, 43.41]
ChainFlow-VLA	35.13	[31.19, 38.83]
DrivoR	34.78	[30.76, 38.58]
ReCogDrive-IL	34.31	[30.38, 38.01]
SUV	32.03	[27.10, 36.55]
ReCogDrive-RL	31.64	[27.82, 35.18]
GoalFlow	31.54	[27.26, 35.45]
NoRD	29.51	[25.79, 33.03]
WoTE	28.52	[24.28, 32.28]
AutoVLA	27.08	[22.93, 30.83]
TransFuser	27.01	[22.48, 31.03]
SparseDriveV2	26.60	[22.79, 30.05]
DriveVLA-W0	16.82	[12.11, 20.95]
DiffusionDrive	14.68	[9.92, 18.69]
StateMLP	11.63	[9.78, 13.37]
DP-VLA	10.90	[6.13, 14.88]
Curious-VLA	6.02	[0.00, 9.72]
CLOVER	5.86	[0.00, 9.45]
WAM-Flow	5.86	[0.00, 9.36]
SimWAM	4.12	[0.00, 7.41]
DiffusionDriveV2	0.00	[0.00, 0.00]
Drive-JEPA	0.00	[0.00, 0.00]
DynVLA Stage 2	0.00	[0.00, 0.00]

DiffusionDriveV2, Drive-JEPA, and DynVLA Stage 2 have NAS = 0, yielding NCS = 0 in both the original evaluation and all bootstrap replicates.

For each bootstrap replicate, 5,911 scene tokens are sampled with replacement. All scene–condition instances associated with a sampled scene are retained together. Predictions from all evaluated models associated with the same sampled scene are also retained together. This produces matched bootstrap samples across models and preserves the paired evaluation structure.

We use 10,000 scene-level bootstrap replicates to estimate uncertainty. Table 9 reports NCS and its 95% percentile bootstrap confidence interval for each model.

Pairwise model comparisons are performed using the same matched bootstrap replicates. For a pair of models a and b, let

$$\Delta_r^* = \mathrm{NCS}_a^{*(r)} - \mathrm{NCS}_b^{*(r)} \tag{123}$$

denote their NCS difference in bootstrap replicate r.

For $B = 10{,}000$ matched bootstrap replicates, the two-sided p-value is computed using an add-one correction:

$$p = \min\left\{1, 2\min\left(\frac{1+\sum_{r=1}^{B} \mathbb{1}[\Delta_r^* \le 0]}{B+1}, \frac{1+\sum_{r=1}^{B} \mathbb{1}[\Delta_r^* \ge 0]}{B+1}\right)\right\}. \tag{124}$$

Statistical testing is performed over all

$$\binom{23}{2} = 253 \tag{125}$$

pairwise NCS hypotheses. Benjamini–Hochberg correction is applied jointly to these 253 p-values to control the false discovery rate.

We assess uncertainty in model comparisons using matched scene-level bootstrap resampling. Of the 253 pairwise NCS comparisons, 193 are significant after Benjamini–Hochberg correction at $q < 0.05$. OneVL exceeds ChainFlow-VLA by 4.17 points, with a 95% confidence interval of [0.74, 7.43] for the paired difference and an adjusted q-value of 0.0252. For adjacent models ranked second through seventh, the 95% confidence intervals for paired NCS differences all include zero.

Bootstrap confidence-interval exclusion and Benjamini–Hochberg significance are reported as two related statistical criteria. The same scene-level resampling principle is used for the confidence intervals reported in the contextual factor analysis.

D.4 Implementation and Reproducibility Details

All reported PlanContext results use the same fixed scenario construction and evaluation rules described in Secs. A, B, and C. The category definitions, geometric thresholds, path-change criteria, GT-relative suppression rules, LC continuity rules, road-compliance constraints, and metric aggregation are shared across all evaluated models.

The evaluation operates on model trajectories after conversion to a common representation of eight future ego poses over four seconds. Predicted poses are interpreted at the rear axle center before geometric evaluation. Road-relative projection, topology assignment, paired trajectory comparison, road-compliance evaluation, and NAVSIM safety evaluation are applied using the same procedure for every model.

Cases with ambiguous scene geometry are removed during benchmark construction according to Sec. A. Geometry-related states produced during model evaluation are handled according to the rules in Sec. B. For category LC, predictions assigned the *uncertain* continuity state are excluded from the valid denominator when computing the category success rate.

The benchmark construction and evaluation pipeline uses fixed category assignments and fixed evaluation thresholds for all reported results. Model comparisons are performed on the same scene–condition definitions under the same evaluation protocol.

D.5 Context Factor Definitions

The contextual factor analysis examines how in-context driving performance varies with scene structure and traffic interaction. All factor stratification is performed within each PlanContext category. This avoids directly mixing the different scene distributions of the six contextual categories.

Decision Urgency. Decision urgency characterizes how close the current scene is to requiring an immediate behavioral decision. Since the six categories represent different contextual requirements, the corresponding proxy is category specific:

$$u_i^{\mathrm{urg}} = \begin{cases} 1/t_{\mathrm{arrival}}^{\mathrm{robust}}, & k_i \in \{A, B1\}, \\ 1/d_{\mathrm{commit}}, & k_i = C, \\ 1/L_{\mathrm{parallel}}^{\mathrm{remain}}, & k_i = L, \\ M_{\mathrm{lat}}^{\mathrm{hist}}, & k_i = LC, \\ D_{\mathrm{GT}}^{4s}/d_{\mathrm{decision}}, & k_i = B2. \end{cases} \tag{126}$$

Here, $L_{\mathrm{parallel}}^{\mathrm{remain}}$ denotes the remaining parallel length of the target adjacent lane. $M_{\mathrm{lat}}^{\mathrm{hist}}$ denotes the magnitude of historical lateral motion used to characterize the ongoing maneuver. Larger values indicate higher decision urgency.

Within each category, urgency values are ranked and divided into Low, Medium, and High groups using category-specific rank terciles.

Geometric Salience. Geometric salience characterizes how clearly the local geometry exposes the relevant decision structure. It is constructed from the available normalized geometric components

$$\frac{|\theta_{\mathrm{start}}|}{15^\circ}, \qquad \frac{|\theta_{\mathrm{branch}}^{\max}|}{45^\circ}, \qquad \frac{d_{\mathrm{split}}}{5\ \mathrm{m}}, \tag{127}$$

$$\frac{|\Delta\theta_{\mathrm{hist}}|}{30^\circ}, \qquad \frac{|\Delta\theta_{\mathrm{lane}}|}{15^\circ}, \qquad \frac{d_{\mathrm{lateral}}}{8\ \mathrm{m}}. \tag{128}$$

Each available component is clipped to the interval $[0, 2]$. The geometric salience score is defined as

$$s_i^{\text{geo}} = \frac{1}{|\mathcal{V}_i|} \sum_{m \in \mathcal{V}_i} \text{clip}\left(g_{i,m}, 0, 2\right), \tag{129}$$

where $\mathcal{V}_i$ denotes the set of available geometric components for instance i.

Geometric salience is therefore a composite descriptor built from the available road and motion geometry. Low, Medium, and High groups are obtained using rank terciles within each category.

Topological Evidence. Topological evidence characterizes the strength of the local structural cues associated with the contextual requirement. Its definition depends on the category.

For A and B1, the evidence is based on the relative heading of the target branch. Weak, medium, and strong evidence correspond to

$$|\Delta\theta_{\text{target}}| < 60^\circ, \tag{130}$$

$$60^\circ \leq |\Delta\theta_{\text{target}}| < 90^\circ, \tag{131}$$

and

$$|\Delta\theta_{\text{target}}| \geq 90^\circ. \tag{132}$$

For B2, topological evidence is defined by the number of locally supported targets in the alternative direction. Weak evidence corresponds to at least two supported targets. Medium evidence corresponds to one supported target. Strong evidence corresponds to no supported target. Evidence strength therefore increases as the number of supported alternatives decreases.

For C, the evidence is based on the post-commitment progress. Weak evidence satisfies

$$d_{\text{commit}} < 10 \text{ m}. \tag{133}$$

Medium evidence satisfies

$$10 \text{ m} \leq d_{\text{commit}} < 30 \text{ m}. \tag{134}$$

Strong evidence satisfies

$$d_{\text{commit}} \geq 30 \text{ m}. \tag{135}$$

For L, the evidence is determined by the available parallel length of the target adjacent lane. Weak, medium, and strong evidence correspond to 20–30 m, 30–60 m, and at least 60 m, respectively.

For LC, the evidence is based on historical lateral-motion magnitude. Weak evidence corresponds to less than 1.0 m. Medium evidence corresponds to 1.0–1.5 m. Strong evidence corresponds to at least 1.5 m.

The topological evidence analysis in the main paper reports Strong-minus-Weak contextual success.

Traffic Density. Traffic density is defined using a weighted count of nearby traffic participants within 40 m:

$$D_{\text{traffic}} = N_{\text{veh},40m} + 0.5\left(N_{\text{ped},40m} + N_{\text{bike},40m}\right) + 1.5N_{\text{front},40m}. \tag{136}$$

Here, $N_{\text{veh},40m}$, $N_{\text{ped},40m}$, and $N_{\text{bike},40m}$ denote the numbers of vehicles, pedestrians, and bicycles within 40 m of the ego vehicle. $N_{\text{front},40m}$ denotes the number of traffic participants located within the forward driving corridor.

Traffic density is divided into Low, Medium, and High groups using category-specific rank terciles.

D.6 Traffic Interaction Comparisons

The traffic-interaction analysis compares contextual success between predefined comparison and reference conditions within the same PlanContext category.

For Lead Vehicle, the comparison condition is ≤ 10 m and the reference condition is > 25 m. For TTC Risk, the comparison condition is < 3 s and the reference condition is ≥ 6 s.

Table 10: Robustness of the PlanContext ranking under metric variations.

Setting	Minimum Spearman correlation	Maximum rank change
$w \in [0.40, 0.60]$	0.9960	1
$w \in [0.30, 0.70]$	0.9703	5
Omit B1 from CPS	0.9960	2
Omit B2 from CPS	0.9970	1
Omit C from CPS	0.9970	1

For Target Occupancy and Crossing Conflict, the comparison condition contains at least two actors and the reference condition contains no actors. Traffic Density compares the High group with the Low group.

All reported differences are computed as

$$\Delta q = q_{\text{comparison}} - q_{\text{reference}}. \tag{137}$$

For Traffic Density, contextual success is pooled over valid model–instance predictions separately within the High and Low groups. We draw 10,000 bootstrap samples from the unique scene tokens within each category, sampling with replacement and retaining all associated scene–condition instances and model predictions together. Each replicate yields a High-minus-Low difference, from which we compute a 95% percentile confidence interval. Asterisks in Table 5 indicate intervals that exclude zero.

The reported differences describe within-category associations between contextual factors and model performance. They do not assume causal effects of the corresponding traffic conditions.

D.7 Metric Robustness Analysis

We evaluate the sensitivity of model rankings to the relative weighting of NAS and CPS using

$$\text{NCS}(w) = \text{NAS}^{w}\text{CPS}^{1-w}. \tag{138}$$

We vary w from 0.30 to 0.70 in increments of 0.01 and compare each ranking with the default setting of $w = 0.5$. Within [0.40, 0.60], the minimum Spearman correlation is 0.9960 and the maximum rank change is one position. Over [0.30, 0.70], these values are 0.9703 and five positions, respectively. OneVL remains the highest-ranked model at every tested weight.

We further examine sensitivity to the composition of CPS by omitting B1, B2, or C in turn and averaging the two remaining category scores. NAS remains fixed in each comparison. The resulting rankings have Spearman correlations of 0.9960, 0.9970, and 0.9970 with the default ranking, with maximum rank changes of two, one, and one positions, respectively.

D.8 Atomic Metric Decomposition

We further decompose the final contextual success into its underlying evaluation components. This analysis helps identify whether failures arise from alternative-condition correctness, road compliance, dynamic safety, or the paired contextual response.

Table 11 reports the category-level decomposition pooled over all evaluable model-instance predictions. The pooled LC success rate is 85.64%, compared with the model-averaged rate of 85.51% in Table 3, because the number of evaluable LC predictions varies across models. Alt.-Condition Correctness denotes the category-specific correctness under the alternative navigation condition. It is topology-based for A and L, includes GT-relative suppression for B1, B2, and C, and corresponds to continuity correctness for LC. RoadOK denotes the full-vehicle road-compliance indicator R_i^F. Dynamic Safety denotes the NAVSIM safety indicator E_i^F.

Overall Safety requires both road compliance and dynamic safety:

$$\text{OverallSafety}_i = R_i^F \wedge E_i^F. \tag{139}$$

Alternative-Condition Success additionally requires the category-specific alternative-condition correctness A_i^F defined in Sec. C.3:

$$q_i^{\text{alt}} = A_i^F \wedge R_i^F \wedge E_i^F . \tag{140}$$

Contextual Success further incorporates the paired contextual decision correctness C_i:

$$q_i = C_i \wedge R_i^F \wedge E_i^F . \tag{141}$$

Table 11: Atomic metric decomposition across the six PlanContext categories.

Category	Alt. Correct.	RoadOK	Dynamic Safety	Overall Safety	Alt. Success	Contextual Success
A	44.94	19.49	66.12	11.95	11.44	10.01
B1	94.37	68.63	85.43	63.59	62.91	62.91
B2	92.50	79.86	84.19	73.95	73.28	73.28
C	88.12	46.87	67.93	35.85	35.01	34.98
L	3.51	97.57	94.16	93.55	1.58	0.45
LC	97.51	94.32	87.05	85.85	85.64	85.64

• Values are pooled over all evaluable model-instance predictions within each category.

The decomposition reveals distinct failure profiles across contextual requirements. Category A is constrained by both alternative-condition correctness and road compliance. Its Alt.-Condition Correctness is 44.94%, while RoadOK is only 19.49%, resulting in a contextual success rate of 10.01%. Category L shows a different pattern. RoadOK and Dynamic Safety reach 97.57% and 94.16%, respectively, whereas Alt.-Condition Correctness is only 3.51%. This suggests that failures in L primarily arise from the inability to execute the feasible alternative lane intention.

Category C retains relatively high alternative-condition correctness at 88.12%, while RoadOK and Dynamic Safety reduce the final contextual success to 34.98%. B2 and LC achieve the highest contextual success among the six categories at 73.28% and 85.64%, respectively. For B1, the high Alt.-Condition Correctness does not imply that every detected behavioral response is appropriate because longitudinal adjustment and waiting can remain valid under its contextual requirement.

D.9 Human Expert Validation

After the scenario construction and evaluation rules were finalized, three reviewers with backgrounds in autonomous driving independently inspected the same randomly sampled subset of 500 scene–condition instances and 11,500 model-instance evaluations from all 23 evaluated models. The reviewers were blinded to the automated evaluation results during independent assessment. They subsequently discussed their judgments to reach a final consensus.

For scenario construction, the reviewers identified 3 of the 500 scene–condition instances whose assigned contextual categories were inconsistent with their consensus judgments. For model-instance evaluation, the final consensus differed from the automated evaluation outcomes for approximately 0.7% of the 11,500 inspected model-instance evaluations.

The audit also identified ambiguity in the original NAVSIM navigation labels. Further examination of their construction provides context for this observation. NAVSIM assigns directional commands by thresholding the lateral displacement of a point 20 m ahead along the route centerline, with left/right thresholds of ± 2 m (Dauner et al., 2024). Its left/right commands encompass turns, lane changes, and sharp curves, leaving the maneuver type unspecified. In the inspected cases, some lane changes were associated with left/right labels, while others were not.

This ambiguity may partly contribute to the poor performance observed in category L, since the same left/right commands can correspond to different maneuver types during training.

E QUALITATIVE EXAMPLES

E.1 EXAMPLES ACROSS CONTEXTUAL CATEGORIES

Figures 4–7 present qualitative examples from categories A, B1, B2, and C to illustrate model responses to navigation interventions under different driving contexts. Each figure shows the shared camera observations and trajectories predicted under the original and alternative navigation commands. The displayed scores summarize model performance over the corresponding evaluation sets.

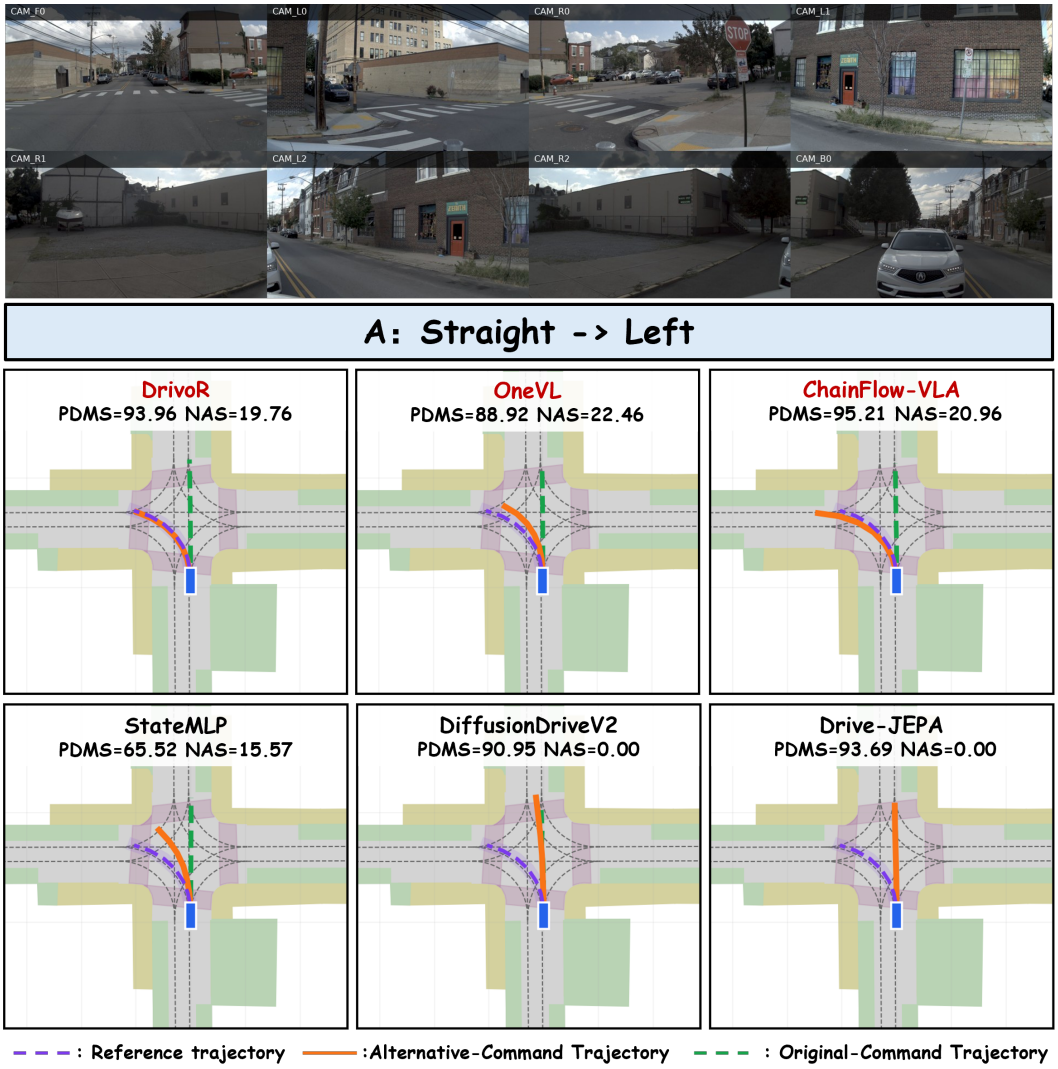


Figure 4: Category A: an immediately actionable alternative route. The navigation command changes from straight to left at an intersection where the alternative left-turn route is reachable within the evaluation horizon. The trajectories illustrate differences in how models respond to the alternative navigation intention.

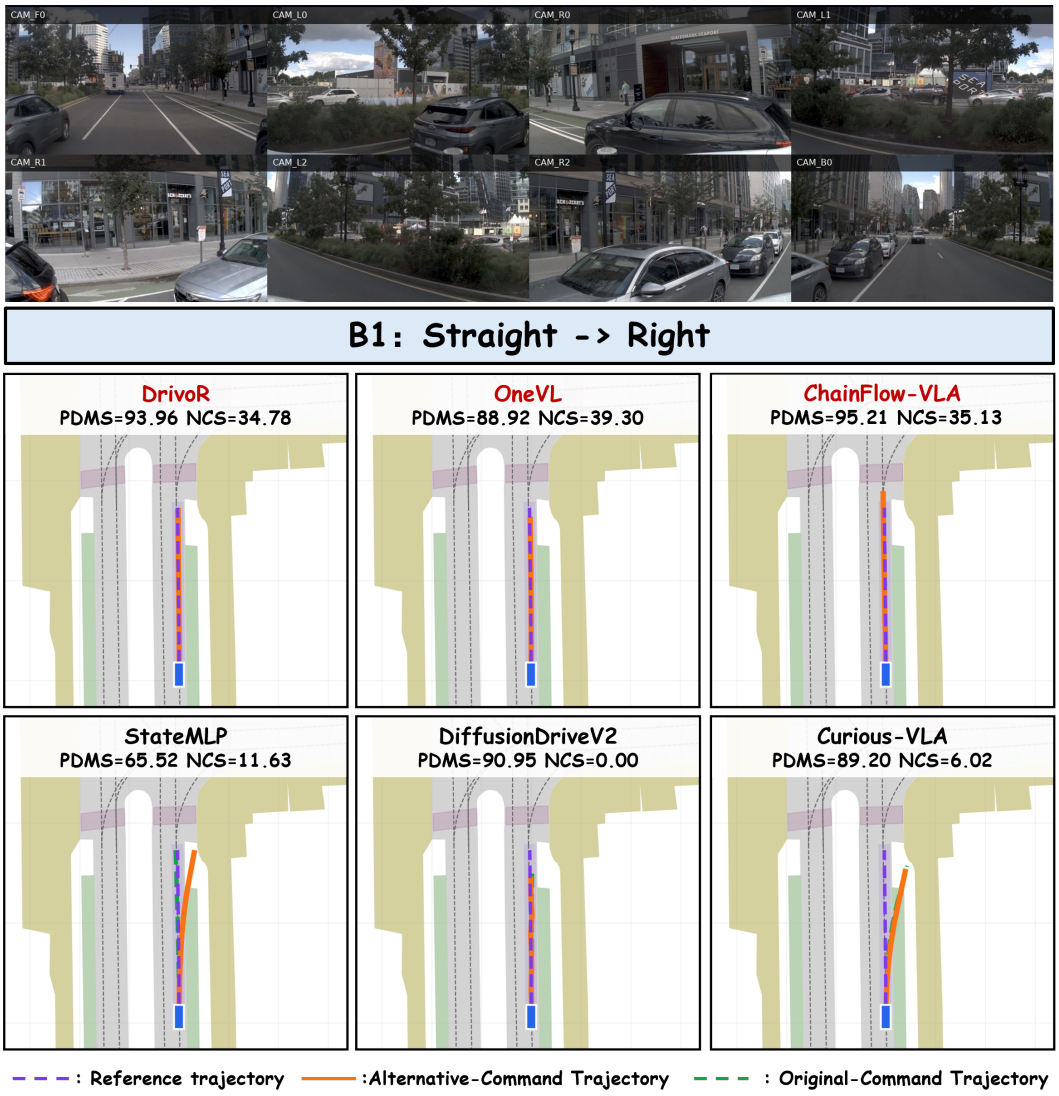


Figure 5: Category B1: an alternative route beyond the current decision window. The navigation command changes from straight to right before the right-turn target becomes actionable. Continuing along the current road remains appropriate within the evaluation horizon. The trajectories illustrate differences between maintaining the current path and deviating prematurely toward the requested direction.

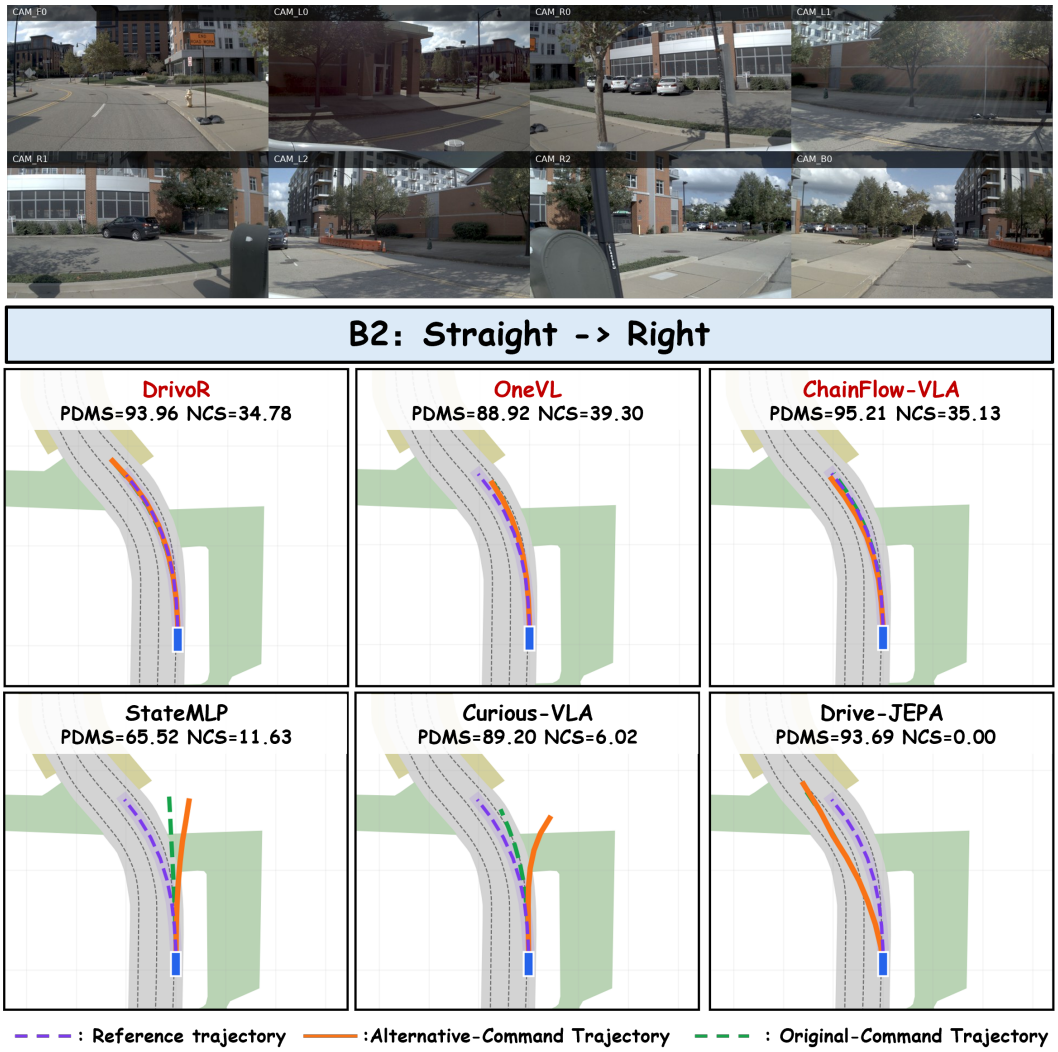


Figure 6: Category B2: a locally unsupported navigation intention. The navigation command changes from straight to right where the local road topology provides no feasible right-turn route. The trajectories illustrate differences between following the existing road geometry and deviating toward the unsupported direction.

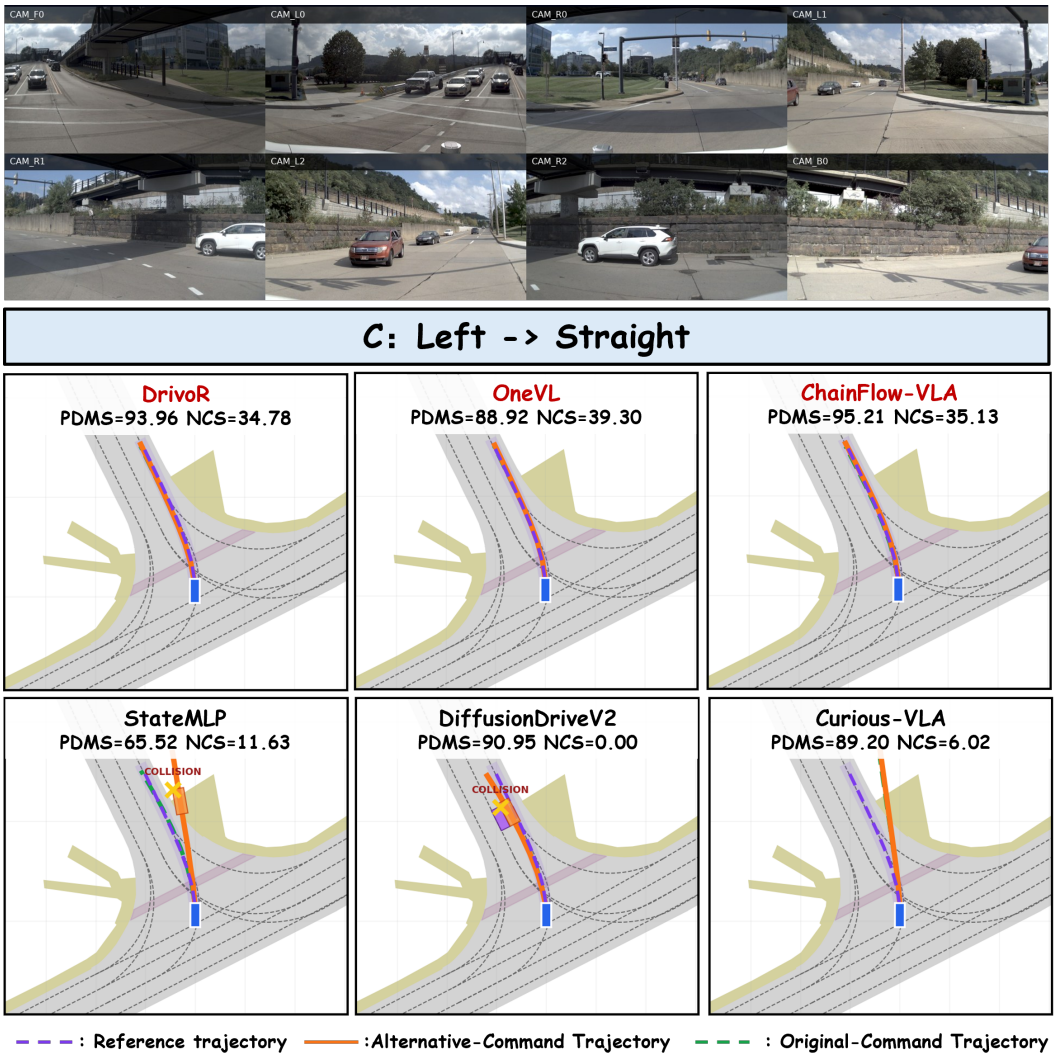


Figure 7: Category C: a navigation intervention after route commitment. The navigation command changes from left to straight after the ego vehicle has passed the route commitment point. The trajectories illustrate differences between continuing along the committed route and deviating in response to the changed command. Collision annotations identify the marked failures in the displayed predictions.

E.2 Examples across Models

We present additional qualitative examples for ten evaluated models in Figures 8–17. Each figure contains three selected scene–condition instances, with camera views and corresponding bird's-eye-view visualizations of the predicted trajectories. Green denotes predictions under the original navigation command, while yellow in the camera views and orange in the bird's-eye views denote predictions under the alternative command. The examples illustrate route choices, responses to navigation interventions, and safety violations across different driving contexts. The displayed mPDMS and NCS values summarize model performance over the corresponding evaluation sets.

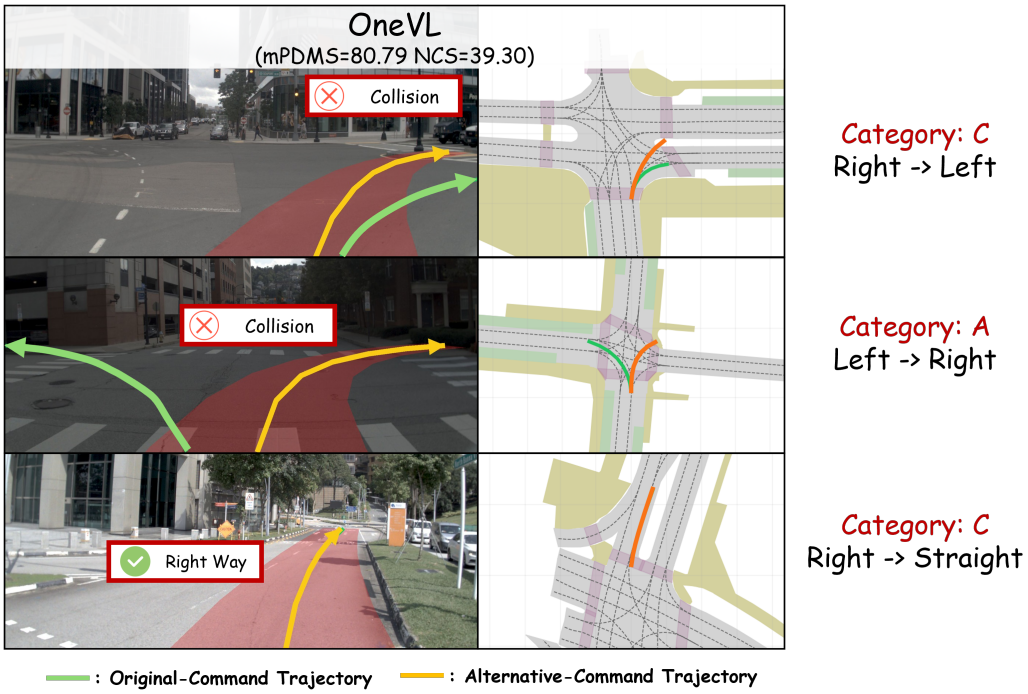


Figure 8: Qualitative examples for OneVL. The first two cases show collisions following navigation interventions in categories C and A. The third case shows a directionally correct response in category C.

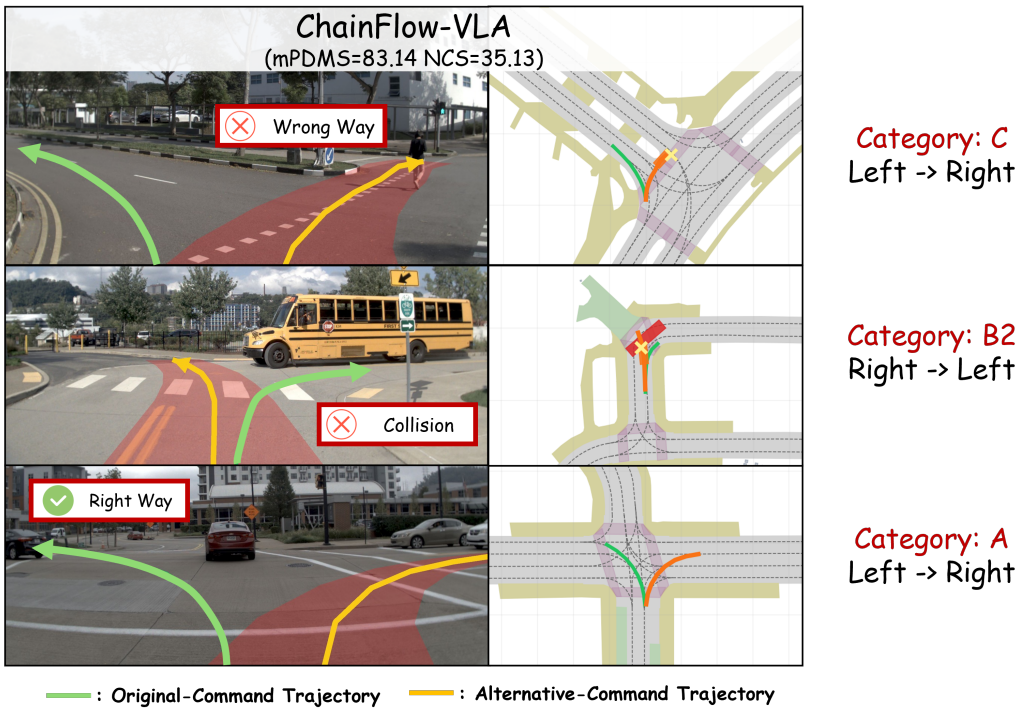


Figure 9: Qualitative examples for ChainFlow-VLA. The cases show an incorrect route choice after route commitment in category C, a collision in category B2, and a correct alternative route choice in category A.

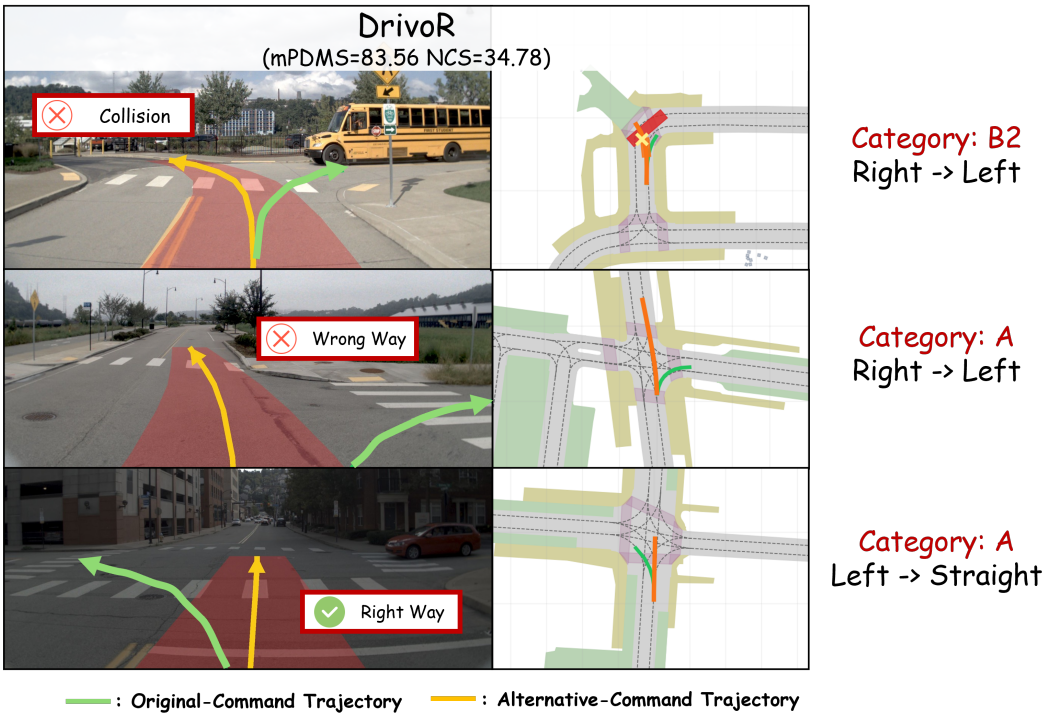


Figure 10: Qualitative examples for DrivoR. The first case shows a collision in category B2. The two category A cases illustrate different outcomes when responding to an immediately actionable alternative navigation intention, including an incorrect route choice and a correct route choice.

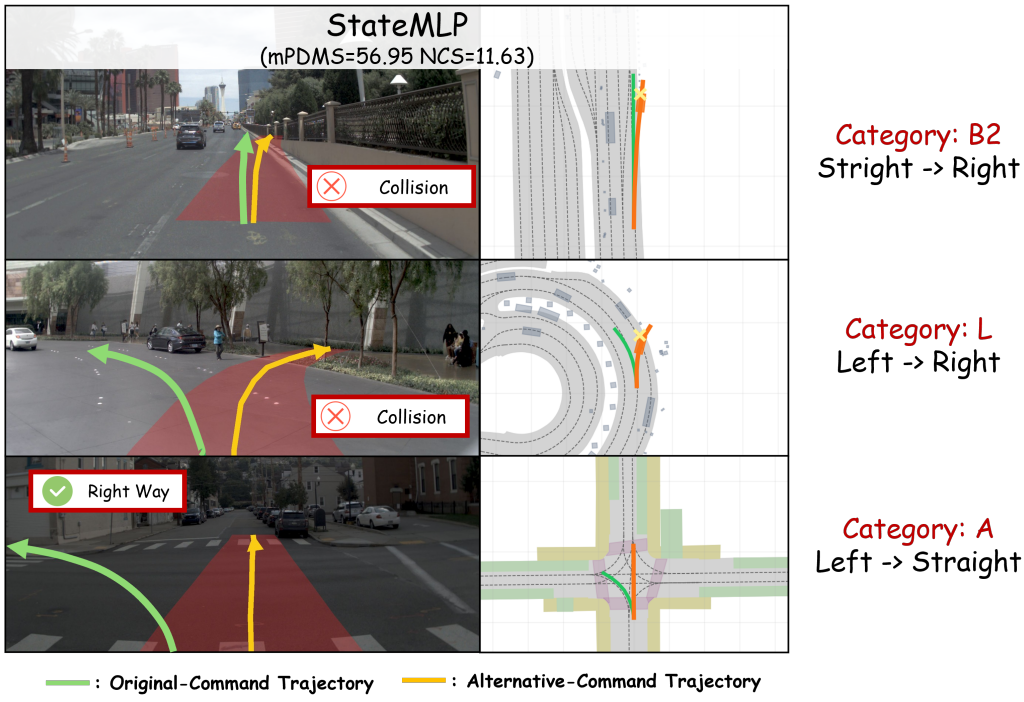


Figure 11: Qualitative examples for StateMLP. The first two cases show collisions in categories B2 and L. The third case shows a correct alternative route choice in category A.

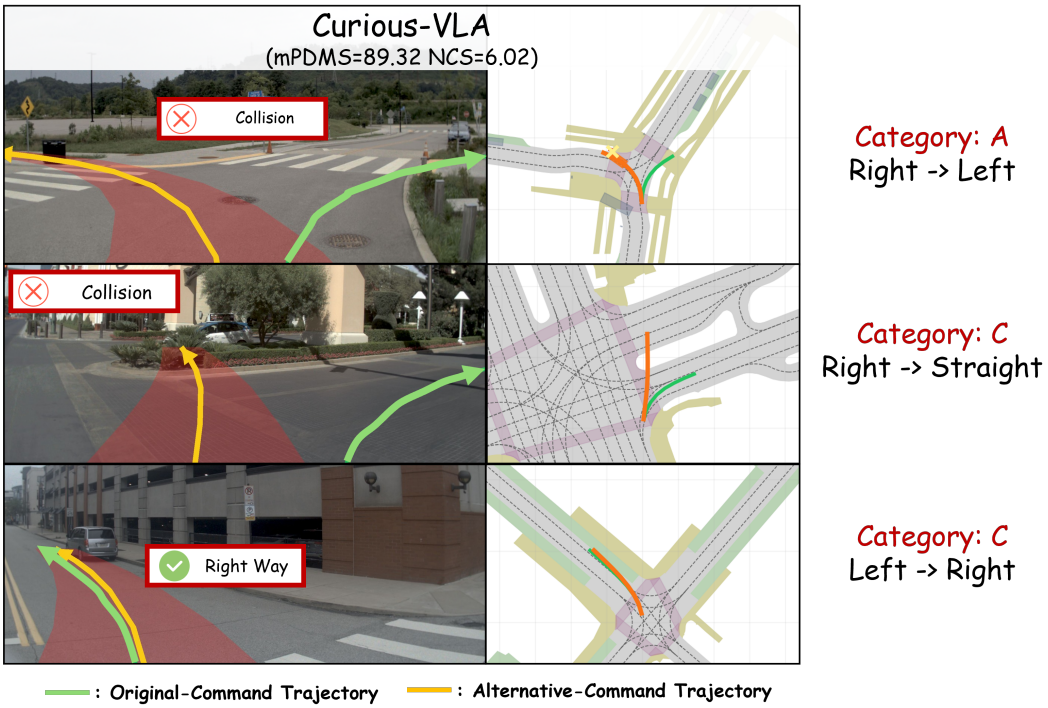


Figure 12: Qualitative examples for Curious-VLA. The first two cases show collisions in categories A and C. The third case shows continued travel along the committed route after the navigation command changes in category C.

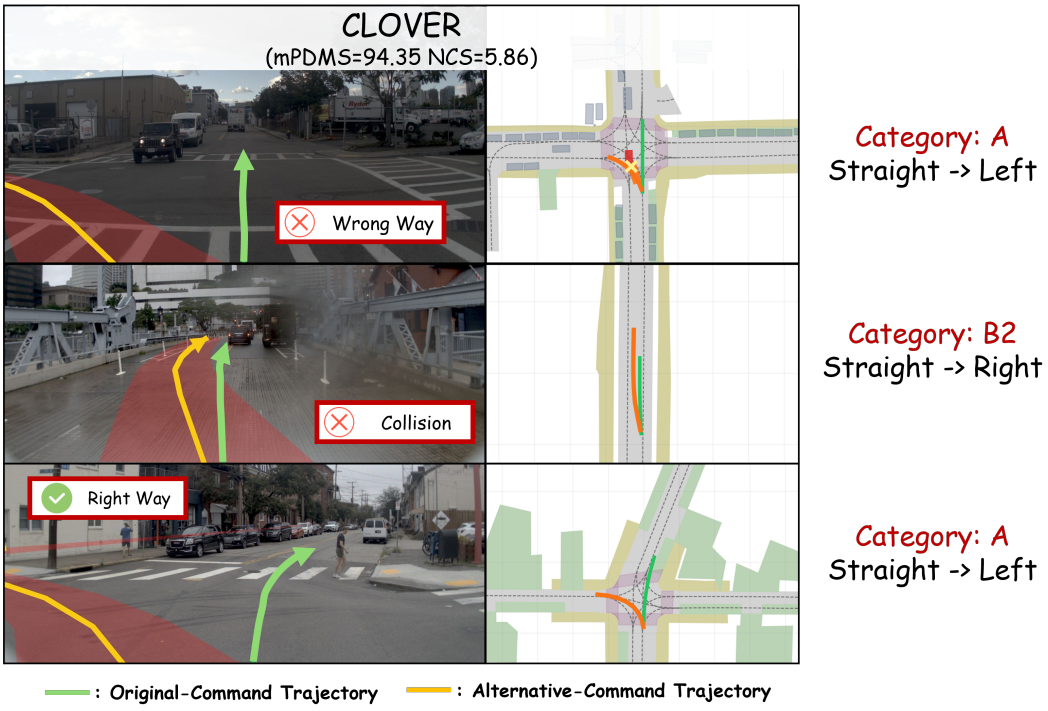


Figure 13: Qualitative examples for CLOVER. The cases show an incorrect alternative route choice in category A, a collision in category B2, and a correct alternative route choice in another category A scene.

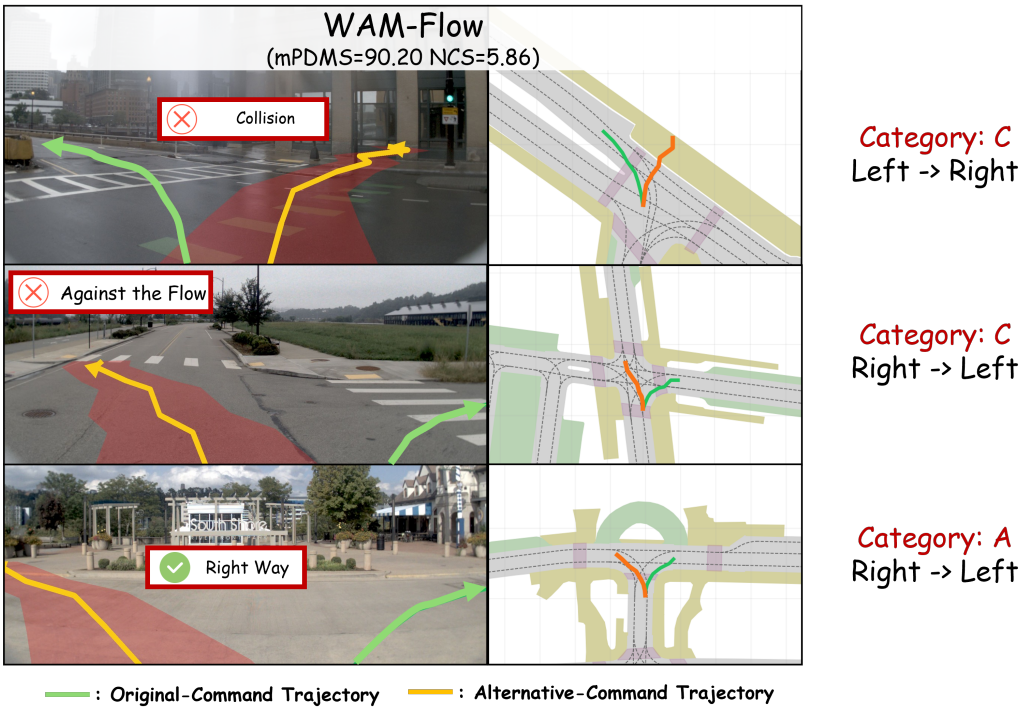


Figure 14: Qualitative examples for WAM-Flow. The two category C cases show a collision and travel against the traffic direction following navigation interventions after route commitment. The category A case shows a correct alternative route choice.

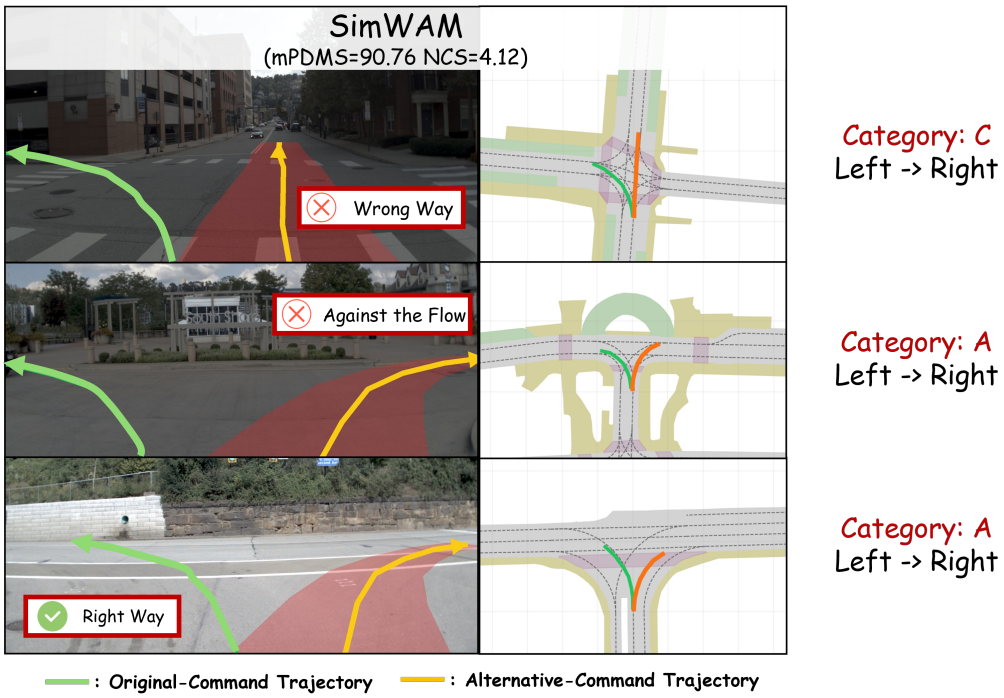


Figure 15: Qualitative examples for SimWAM. The first case shows an incorrect route choice in category C. The two category A cases show travel against the traffic direction and a correct alternative route choice, respectively.

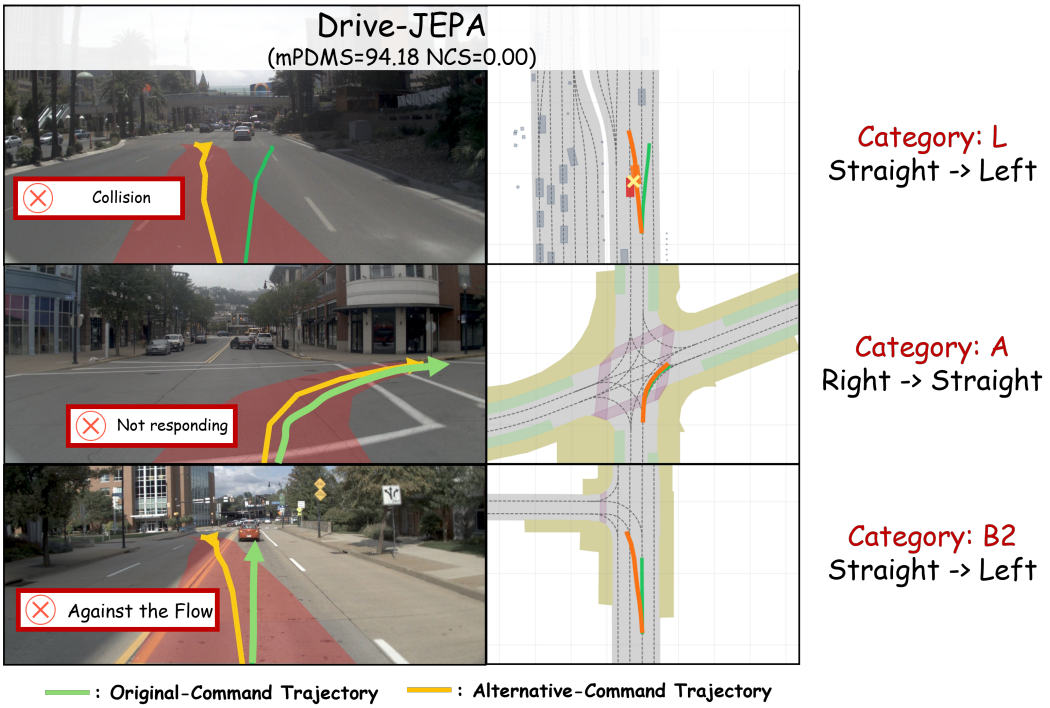


Figure 16: Qualitative examples for Drive-JEPA. The cases show a collision in category L, little visible change between the paired predictions in category A, and travel against the traffic direction in category B2.

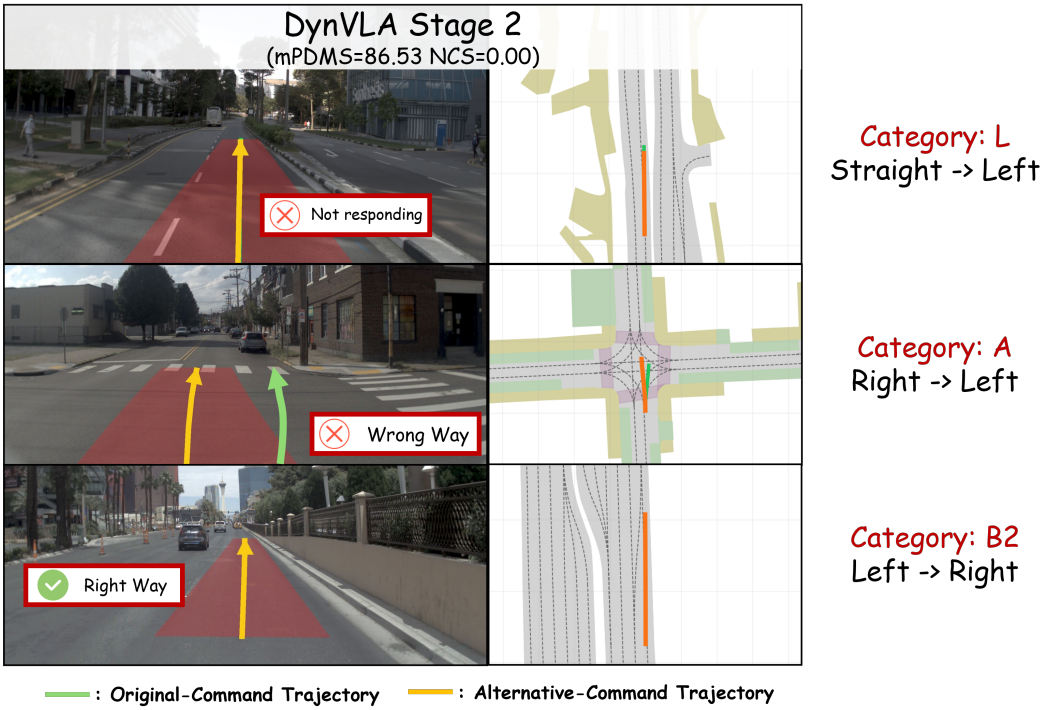


Figure 17: Qualitative examples for DynVLA Stage 2. The cases show little visible response to a lane-level navigation intervention in category L, an incorrect route choice in category A, and continued travel along the current road in category B2.