WHEN THE SCORE BECOMES THE TARGET: RETHINKING METRIC VALIDITY IN AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Driving benchmark scores are increasingly used not only for evaluation but also as optimization targets. This raises a fundamental question: do score gains remain reliable evidence of driving improvement once the score itself is optimized? We address this question by examining how the scoring process responds to changes in driving behavior and whether the resulting gains persist under repeated execution and replanning. We decompose the process into execution, measurement, subscore mapping, and aggregation. Controlled interventions reveal substantial behavioral changes that receive little score response because distinctions are omitted, thresholded, or attenuated between requested and executed motion. Closed-loop comparisons further show that optimization gains can reverse when the execution interface changes, demonstrating their dependence on how requests are executed and returned as feedback. Together, these findings connect the behavioral distinctions preserved by a metric to the conditions under which its gains transfer. Metric validity under optimization therefore requires examining both what the scoring process measures and how the optimized behavior is executed.

1 INTRODUCTION

Autonomous driving aims to transport people safely, comfortably, and efficiently to their destinations, while following traffic rules and interacting safely with other road users.

These goals motivate driving benchmarks (Caesar et al., 2021; Dauner et al., 2024; Jia et al., 2024; Zhou et al., 2024) and learned planning systems (Liao et al., 2025; Zheng et al., 2025; Li et al., 2026b). Benchmark scores provide a common basis for comparing progress toward these goals and increasingly serve as rewards or trajectory-selection objectives (Rawal et al., 2026; Li et al., 2025; Xu et al., 2026). Once a score becomes an optimization target, a central question is whether its gains continue to reflect improvements in the driving behaviors it is intended to measure, a property we refer to as *metric validity*. Optimization changes the trajectories a planner produces, so a score's ability to distinguish behavior on existing outputs does not establish its validity on optimized outputs (Figure 1). This concern is related to Goodhart's law (Manheim & Garrabrant, 2018; Karwowski et al., 2024), while prior driving studies have exposed evaluation shortcuts and discrepancies between open- and closed-loop performance (Zhai et al., 2023; Dauner et al., 2023; Wang et al., 2026). Recent post-training results further show that open-loop gains do not always translate into closed-loop improvements (Li et al., 2026a). These findings establish that score–behavior discrepancies can occur, but leave open where they arise in the driving evaluation process and how optimization exposes them. We address this gap by examining behavioral distinctions obscured by score aggregation, thresholding, and transformations between planned and executed motion, and by tracing whether optimization gains persist under repeated replanning. Our analysis and large-scale study provide practical guidance on which behaviors a metric should distinguish and how its validity should be reassessed after optimization.

2 SCORE–BEHAVIOR ANALYSIS

Let S denote the reported benchmark score and D_k a behavior-specific diagnostic relevant to an improvement claim. Let R denote the optimization objective, which may equal S, approximate it, or include auxiliary terms. An optimization procedure $\mathcal{O}_R$ transforms a planner π_0 into $\pi_R = \mathcal{O}_R(\pi_0)$

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.

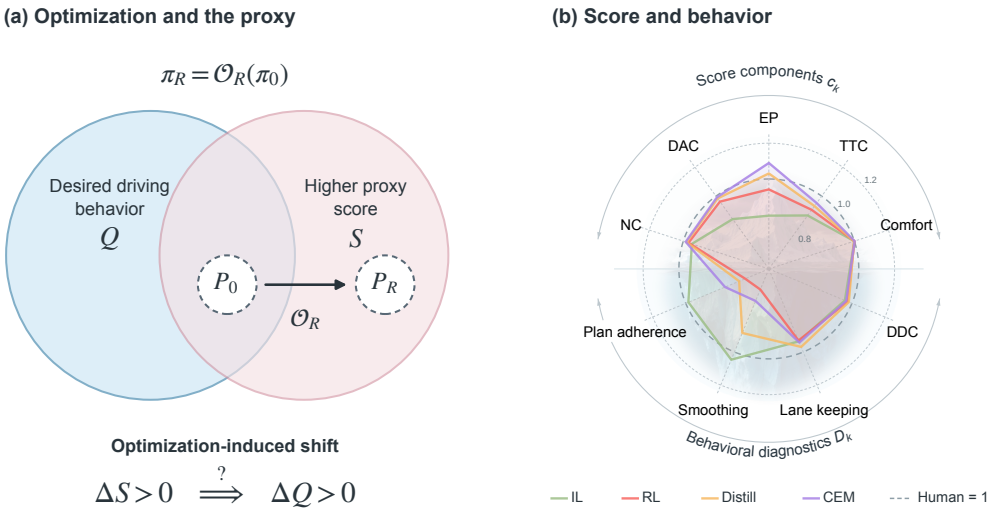

Figure 1: **Optimization and score–behavior mismatch.** (a) Conceptual distributions: optimizing R changes P_0 to P_R; higher S does not imply better Q. (b) Group-mean score components c_k and diagnostics D_k for PDMS (Predictive Driver Model Score). Each axis is oriented so that a larger radius represents better compliance or a smaller behavioral error. Values are normalized relative to the logged-human mean, shown at radius 1, to place heterogeneous quantities on a common visual scale. The radial scale is nonlinear.

through training or trajectory selection, inducing output distributions $P_{0,e}$ and $P_{R,e}$ under evaluation setting e on a shared initial-scene distribution. For any measured quantity F, we define

$$\Delta F_e = \mathbb{E}_{P_{R,e}}[F_e] - \mathbb{E}_{P_{0,e}}[F_e]. \tag{1}$$

Our central question is whether a score gain, $\Delta S_e > 0$, is accompanied by improvement in the driving behaviors the score is intended to measure. We assess this correspondence using diagnostics D_k specific to the behavior of interest, rather than reducing driving quality to a single surrogate.

Metric validity therefore depends on both the outputs and how they are evaluated: validity on $P_{0,e}$ does not by itself establish validity on $P_{R,e}$ or under a different execution or feedback setting. For a request u at a fixed initial state, with execution T_e, measurements m_e, subscore mappings h_e, and aggregation G_e, the score is

$$\bar{S}_e(u) = G_e(h_e(m_e(T_e(u)))) \,. \tag{2}$$

This decomposition identifies three possible blind spots: omission, insensitive subscore mappings, and transformations between requested and scored motion. For each, Section 3 tests the predicted insensitivity and checks whether the corresponding behavior deteriorates in optimized outputs despite a score gain.

Aggregation and omission. Suppressing e, represent a measured behavior by a candidate component $c_k = h_k(D_k)$ and write $S = G(c_1, \ldots, c_K)$. On differentiable branches, holding other inputs fixed,

$$\frac{\partial S}{\partial D_k} = \underbrace{\frac{\partial G}{\partial c_k}}_{\text{aggregation sensitivity}} \underbrace{h_k'(D_k)}_{\text{subscore sensitivity}} . \tag{3}$$

Omission makes the first factor zero throughout the relevant range: changing this component alone cannot change S. Matched trajectory probes test this invariance while controlling other scored properties (Section 3.2). Even for included components, an aggregate gain can reflect trade-offs or changes in gate pass rates; component and conditional measurements distinguish these contributions (Section 3.4).

Thresholding and saturation. An included component can instead have zero subscore sensitivity. For an upper-bounded quantity d,

$$c(d) = \mathbf{1}[d < b], \qquad c'(d) = 0 \quad \text{for } d < b. \tag{4}$$

Values inside the passing region have equal scores when other inputs are fixed; clipping has analogous flat regions. Continuous readouts reveal changes within them, while rescoring fixed outputs with a different h_e tests score resolution (Section 3.3); finite differences handle boundaries.

Execution transformation. Let $s_e = G_e \circ h_e \circ m_e$ score executed states, and let $u + v$ be a finite edit of the request. Its score response is

$$\delta_v \bar{S}_e(u) = s_e(T_e(u + v)) - s_e(T_e(u)) \,. \tag{5}$$

A substantial change in requested motion need not produce a substantial score response: T_e can attenuate it before measurement, and h_e can flatten the remaining response. Section 3.4 compares requested and executed motion and measures Eq. 5 for smoothing and retiming interventions. These checks locate weak sensitivity in the complete composition; they do not separately identify each stage's contribution.

Closed-loop execution and gain transfer. Actions change subsequent planner inputs (Ross et al., 2011). Let z_t contain observations, motion state, and history; Φ_e includes execution, state extraction, rendering, and history update. For a planner π,

$$u_t = \pi(z_t), \qquad z_{t+1} = \Phi_e(z_t, u_t). \tag{6}$$

Changing execution can alter subsequent requests even when initial predictions match. For $J_e(\pi) = \mathbb{E}_{P_{\pi,e}}[S_e]$, compare both planners under settings a and b on shared initial scenes:

$$I_{a,b} = \Delta S_a - \Delta S_b = \big[J_a(\pi_R) - J_b(\pi_R)\big] - \big[J_a(\pi_0) - J_b(\pi_0)\big]. \tag{7}$$

Section 4 measures this interaction for two execution interfaces within otherwise fixed closed-loop conditions. A reversal requires opposite signs of ΔS_a and ΔS_b; $I_{a,b} \neq 0$ alone does not establish one. Comparing successive plans at the same future time then examines how the changed execution and feedback process affects later requests. Together, these measurements test gain transfer and describe the accompanying behavioral response, without assigning the reversal to a particular filtered motion component.

Score insensitivity permits a mismatch; R and planner constraints also shape behavior. We therefore examine controlled interventions and optimized outputs separately, including cases where score gains and behavioral improvements agree.

3 Behavioral Sensitivity of the Scoring Process

3.1 Experimental Setup

Scores and targets. NAVSIM aggregates simulated subscores in $[0, 1]$ using the original PDM Score notation (Dauner et al., 2024):

$$\text{PDMS} = \underbrace{\left(\prod_{m \in \{\texttt{NC},\texttt{DAC}\}} \texttt{score}_m \right)}_{\text{penalties}} \times \underbrace{\left(\frac{\sum_{w \in \{\texttt{EP},\texttt{TTC},\texttt{C}\}} \texttt{weight}_w \times \texttt{score}_w}{\sum_{w \in \{\texttt{EP},\texttt{TTC},\texttt{C}\}} \texttt{weight}_w} \right)}_{\text{weighted average}}. \tag{8}$$

NC/DAC penalize at-fault collisions/road departures; EP/TTC/C measure progress, time-to-collision, and comfort with weights $5/5/2$. NC assigns 0.5 for static-object collisions. DDC has zero aggregation weight. We use official scoring. NAVSIM-v2 introduces EPDMS, with driving-direction compliance as a gate and additional lane-keeping and comfort components (Cao et al., 2025). We distinguish rescoring fixed outputs with EPDMS from the official two-stage protocol, which requires new predictions at original and synthetic states.

Planner comparisons. ReCogDrive-2B/8B, Qwen-Drive, and NoRD provide within-architecture imitation-to-reinforcement comparisons (Li et al., 2026b; Zhou et al., 2026; Rawal et al., 2026). ReCogDrive and NoRD use PDMS rewards; NoRD retains a fixed human-derived trajectory vocabulary. Qwen-Drive uses EP/TTC/comfort weights $6/4/2$ and a displacement reward toward the logged trajectory. GTRS-Dense changes expert supervision to reward-score distillation (Li et al.,

2025); TOAD adds fixed-weight search against predicted PDMS with comfort and anchoring penalties (Xu et al., 2026). DiffusionDrive v1 (IL)→v2 (RL) evaluates the released version upgrade (Liao et al., 2025; Zou et al., 2025). Released checkpoints are evaluated without retraining; Table 1 reports absolute results and paired changes.

Evaluation. NAVSIM open-loop evaluation tracks a four-second plan with LQR (linear quadratic regulator) and a kinematic bicycle model while other agents replay their logs. We use NAVSIM-v1 `navtest`, the `navhard` test set of NAVSIM-v2, and the WorldEngine test set (corner cases identified through failures of its VADv2-based planner on `navtest`) (Li et al., 2026a). WorldEngine supplies updated observations for replanning every 0.5 s over four seconds. Our primary interface comparison uses all scenes with NR traffic, comparing the released trajectory-playback interface against LQR+bicycle execution under the same scoring settings. Earlier playback comparisons with NR/R traffic are reported separately in Appendix D. On nuPlan, Plan-R1 supplies an imitation/RL pair evaluated on `val14` scenarios from logs with non-reactive (NR) and reactive (R) traffic (Tang et al., 2026; Caesar et al., 2021). Its reward changes CLS weights and gates. Paired changes use scene and log bootstrap intervals and Holm-adjusted tests over fixed families. Evaluation protocols, planner configurations, and statistical procedures are detailed in Appendix A.

3.2 Omission: Unscored Direction

Constructing the blind spot. The Predictive Driver Model Score (PDMS) assigns zero aggregation weight to driving-direction compliance (DDC). With the other scoring inputs fixed, $\partial \mathrm{PDMS}/\partial \mathrm{score}_{\mathrm{DDC}} = 0$. We test this invariance by laterally shifting route-following trajectories while keeping their speed profiles fixed and verifying lane direction against the map. The 42 matched pairs across 37 `navhard` scenes pass the no at-fault collision (NC) and drivable-area compliance (DAC) checks, with near-equal progress and identical time-to-collision (TTC) and comfort scores. The wrong-way trajectories travel 8.80 m against the verified direction on average, yet the mean PDMS difference is only 1.9×10^{-5} (Appendix B.3). Matching the other scored properties leaves a direction violation that the aggregate cannot distinguish.

Direction compliance also falls in optimized checkpoints whose aggregate scores improve. Both ReCogDrive reinforcement learning (RL) checkpoints score higher than their imitation learning (IL) counterparts while direction compliance decreases (Table 1). NoRD improves both its score and most measured diagnostics; its lower-scoring imitation baseline leaves greater headroom for improvement. The omitted component can therefore deteriorate without directly lowering the score, but deterioration is not inevitable.

Test-time search exposes the same unconstrained dimension with fixed model weights. In TOAD's cross-entropy method (CEM) search, the highest-predicted-score candidate gains official PDMS while direction compliance and the extended PDM score (EPDMS) fall (Figure 7). TOAD's own iteration ablation also reports EPDMS deterioration with longer search (Xu et al., 2026, Figure 4(c)); our component analysis shows that direction compliance can fall while the v1 score rises. The regularized deployment rule selects trajectories with higher direction compliance than the highest-predicted-score candidate later in the search. Under the released v2 objective, which includes direction compliance, deployed mean compliance increases slightly on both original and synthetic scenes. The unscored direction dimension worsens in both optimized ReCogDrive checkpoints and score-seeking TOAD candidates, while NoRD provides a counterexample.

3.3 Thresholding: Motion within Passing Regions

Passing regions and score ties. The flat regions in Eq. 4 predict that a passing subscore can conceal continuous behavioral changes; we test its resolution through motion diagnostics and fixed-output rescoring. Thresholded comfort leaves continuous motion differences unresolved within the passing region. Across the evaluated NAVSIM configurations, comfort fails in only 0.14% of planner–scene evaluations. For DiffusionDrive v1, v2, and Human on `navhard`, even the largest within-scene threshold utilization averages roughly one-third of the bound (Appendix Figure 4).

At the aggregate level, score ties provide a complementary check of behavioral resolution during trajectory selection. In DiffusionDrive v2's candidate pools, 274 of 450 `navhard` scenes have

Table 1: **NAVSIM `navtest`: scores and behavioral changes.** Absolute values with within-family changes below optimized entries, computed before rounding; green/red denotes improvement/deterioration. EPDMS rescores fixed outputs. DDC^{v2} excludes intersection positions. Jitter: rad; Smooth./Adh.: smoothing displacement/plan-adherence error [m]; LK: lane keeping; HC: history comfort. † marks changes not supported by the paired statistical criteria in Appendix A.4.

		Aggregate scores		EPDMS components			Behavioral diagnostics			
Planner	Variant	PDMS ↑	EPDMS ↑	DDC^{v2} ↑	LK ↑	HC ↑	DDC^{v1} ↑	Jitter ↓	Smooth. ↓	Adh. ↓
Ego-status MLP	IL	0.664	0.645	0.927	0.792	0.983	0.904	0.49	0.086	0.587
TransFuser	IL	0.839	0.814	0.992	0.848	0.983	0.980	0.50	0.092	0.603
Latent TransFuser	IL	0.835	0.810	0.992	0.853	0.983	0.979	0.53	0.091	0.624
ReCogDrive 2B	IL	0.863	0.834	0.993	0.860	0.983	0.980	0.50	0.105	0.575
	RL	0.905 (+0.042)	0.854 (+0.019)	0.982 (−0.011)	0.802 (−0.058)	0.977 (−0.006)	0.967 (−0.013)	1.69 (+1.19)	0.288 (+0.183)	0.893 (+0.319)
ReCogDrive 8B	IL	0.869	0.842	0.994	0.857	0.983	0.980	0.64	0.109	0.583
	RL	0.903 (+0.035)	0.859 (+0.017)	0.989 (−0.005)	0.832 (−0.025)	0.979 (−0.004)	0.974 (−0.006)	1.56 (+0.92)	0.281 (+0.173)	0.982 (+0.399)
Qwen-Drive	IL	0.879	0.854	0.996	0.863	0.982	0.987	0.48	0.116	0.609
	RL	0.905 (+0.026)	0.873 (+0.019)	0.995 (−0.001)	0.851 (−0.012)	0.978 (−0.004)	0.985 (−0.001†)	0.55 (+0.07)	0.212 (+0.097)	0.564 (−0.046)
Diffusion Drive	v1 IL	0.881	0.852	0.994	0.860	0.984	0.981	0.62	0.103	0.584
	v2 RL	0.909 (+0.029)	0.855 (+0.003†)	0.985 (−0.009)	0.840 (−0.020)	0.891 (−0.093)	0.968 (−0.013)	1.99 (+1.37)	0.387 (+0.284)	1.240 (+0.656)
NoRD	IL	0.738	0.715	0.967	0.813	0.957	0.946	0.99	0.157	0.743
	RL	0.846 (+0.108)	0.812 (+0.097)	0.982 (+0.015)	0.827 (+0.014)	0.977 (+0.021)	0.963 (+0.017)	0.78 (−0.22)	0.126 (−0.030)	0.745 (+0.003†)
GTRS Dense	Expert	0.893	0.876	0.994	0.875	0.969	0.985	0.48	0.088	1.162
	Distill	0.903 (+0.010)	0.879 (+0.003†)	0.993 (−0.001†)	0.866 (−0.009)	0.977 (+0.008)	0.981 (−0.004)	0.50 (+0.02†)	0.087 (−0.001)	1.112 (−0.050)
DrivoR / TOAD	Base	0.948	0.893	0.985	0.849	0.967	0.969	0.39	0.127	0.708
	CEM	0.949 (+0.001†)	0.890 (−0.003)	0.982 (−0.003)	0.840 (−0.008)	0.967 (−0.001†)	0.965 (−0.004)	0.25 (−0.14)	0.129 (+0.002)	0.716 (+0.008)

multiple PDMS maximizers with different heading jitter. The aggregate provides no preference among those geometrically different plans.

Changes after optimization. We further examine in nuPlan whether improved closed-loop task scores are accompanied by deterioration in continuous measures of executed motion. We measure the largest magnitude of longitudinal jerk, the time derivative of longitudinal acceleration:

$$D_{\text{jerk}} = \max_t |j_{\parallel}(t)|, \qquad j_{\parallel}(t) = \frac{\mathrm{d}a_{\parallel}(t)}{\mathrm{d}t}. \tag{9}$$

Plan-R1's mean absolute jerk peak increases after RL in both non-reactive and reactive traffic, with both changes passing the paired statistical checks. Its closed-loop score (CLS), road compliance, progress, and speed-limit compliance improve, while planned jitter and plan churn decrease (Table 2). The binary comfort term has no supported mean change in non-reactive traffic and decreases slightly in reactive traffic. The continuous measurement identifies a motion change that the comfort pass rate alone does not describe.

Changing score resolution. A continuous comfort score can distinguish motion within the passing region, but it still evaluates motion after tracking. We replace the binary mapping h_e with three continuous alternatives while keeping the execution transformation T_e and planner outputs fixed. All three preserve DiffusionDrive v2's score advantage on both `navhard` and the WorldEngine test set (Appendix Table 5).

Table 2: **nuPlan `val14`.** Plan-R1 IL→RL under non-reactive (NR) and reactive (R) traffic.

Quantity	Role in CLS	Δ NR	Δ R
CLS ↑	Aggregate	**+0.034**	**+0.067**
Drivable-area compliance ↑	Gate	**+0.042**	**+0.033**
Planned heading jitter ↓	Absent	**−0.235**	**−0.538**
Absolute lon. jerk peak [m/s^3] ↓	Threshold	**+0.136**	**+0.206**
Absolute jerk peak [m/s^3] ↓	Threshold	**+0.068**	**+0.145**
Comfort term ↑	Weighted term	**−0.002**†	**−0.009**

3.4 Execution Transformation: Requested versus Scored Motion

In Section 2, we predict that substantial request changes can receive weak score responses after execution. we test this through motion-layer comparisons and finite trajectory interventions (Eq. 5). Smoothing displacement measures how far waypoints move under a local smoothing operation; heading jitter measures variation in successive heading changes. For eight planned positions $p_1, \ldots, p_N$ spaced 0.5 s apart, one three-point moving average keeps the endpoints fixed and sets $\tilde{p}_i = (p_{i-1} + p_i + p_{i+1})/3$ for internal points. The two diagnostics are

$$D_{\text{smooth}} = \max_i \|\tilde{p}_i - p_i\|_2, \qquad D_{\text{jitter}} = \sum_{i=2}^{N-1} |\theta_{i+1} - 2\theta_i + \theta_{i-1}|, \tag{10}$$

where θ_i is the unwrapped heading of segment $p_{i-1} \to p_i$, with p_0 the ego origin. Smoothing displacement captures both path bending and waypoint spacing; heading jitter is interpreted with path length. Plan-adherence error measures the maximum separation between the interpolated request $p^{\text{plan}}(t)$ and the tracked position $\hat{p}(t)$:

$$D_{\text{adh}} = \max_t \|\hat{p}(t) - p^{\text{plan}}(t)\|_2. \tag{11}$$

Attenuation through tracking. Planned and executed motion show different changes after optimization. For ReCogDrive-2B, the mean planned absolute acceleration peak rises from 1.14 to 2.84 m/s^2 after RL, while the executed longitudinal peak changes from 0.88 to 1.20 m/s^2. The accompanying increase in plan-adherence error records greater separation between the requested and tracked trajectories (Table 1). For Qwen-Drive RL, planned acceleration/deceleration peaks are 1.65/1.88 m/s^2, whereas the simulated peaks are 0.28/0.53 (Table 6). These peaks describe different motion layers and do not directly measure how much of a request is discarded.

Finite interventions. Large changes to requested geometry produce little response from the complete execution–scoring pipeline. Smoothing DiffusionDrive v2's positions while preserving its output headings removes 70–74% of heading jitter on `navhard` and the WorldEngine test set, with mean PDMS changes below 0.01 in magnitude. The illustrative case shows a visibly oscillatory v2 plan scoring above both v1 and Human, although all three pass comfort after tracking (Figure 2).

Qwen-Drive's increased smoothing displacement is concentrated in longitudinal waypoint spacing, accompanying changes in planned speed; its plan-adherence error improves. Retiming along each original path preserves endpoint speeds and total length while reducing the RL plan's peak speed change per unit time by nearly half, while mean PDMS changes by only -0.0002. Both interventions alter requests before tracking, so their score response includes the execution transformation.

Contrasting responses. NoRD improves PDMS while reducing jitter and smoothing displacement; GTRS-Dense improves adherence, with little change in smoothing displacement. TOAD reduces jitter while slightly increasing smoothing displacement and adherence error. Requested-motion diagnostics distinguish plan quality from the controller's ability to accommodate requests; execution scores alone cannot establish both.

Interpreting aggregate gains. The three blind spots also motivate separating behavioral changes from how the aggregate credits them: reported progress can rise through more trajectories passing

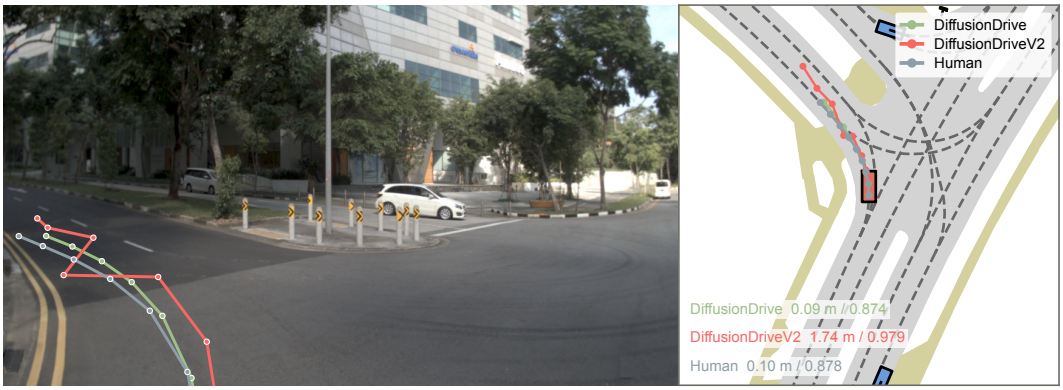

Figure 2: **Illustrative trajectory comparison.** WorldEngine case; labels: smoothing displacement (m, lower is better) / PDMS. The oscillatory v2 plan scores 0.979 versus v1 0.874 and Human 0.878.

Table 3: **NAVSIM-v2 `navhard-two-stage` results.** Selected EPDMS components on original (Stage 1) and synthetic (Stage 2) scenes, and combined score. RL entries show absolute values and parenthesized RL-minus-IL changes. EC: two-frame extended comfort.

		Stage 1 components				Stage 2 components				Overall
Planner	Variant	DAC ↑	DDC ↑	EP ↑	EC ↑	DAC ↑	DDC ↑	EP ↑	EC ↑	EPDMS ↑
ReCogDrive 2B	IL	0.827	0.984	0.828	0.764	0.721	0.823	0.792	0.688	0.278
	RL	0.907 (+0.080)	0.956 (−0.029)	0.856 (+0.028)	0.329 (−0.436)	0.726 (+0.005)	0.812 (−0.011)	0.853 (+0.061)	0.310 (−0.378)	0.270 (−0.008†)
ReCogDrive 8B	IL	0.842	0.990	0.834	0.769	0.721	0.821	0.808	0.680	0.263
	RL	0.911 (+0.069)	0.973 (−0.017)	0.860 (+0.027)	0.204 (−0.564)	0.728 (+0.008)	0.819 (−0.002)	0.864 (+0.056)	0.153 (−0.527)	0.255 (−0.008†)
Qwen-Drive	IL	0.887	0.994	0.829	0.813	0.753	0.843	0.771	0.647	0.303
	RL	0.940 (+0.053)	0.997 (+0.002)	0.839 (+0.010)	0.342 (−0.471)	0.782 (+0.029)	0.851 (+0.008)	0.777 (+0.006)	0.351 (−0.296)	0.306 (+0.003†)
NoRD	IL	0.651	0.923	0.810	0.604	0.593	0.769	0.767	0.525	0.178
	RL	0.836 (+0.184)	0.973 (+0.050)	0.811 (+0.001)	0.653 (+0.049)	0.713 (+0.120)	0.820 (+0.051)	0.808 (+0.040)	0.563 (+0.038)	0.257 (+0.079)

scoring gates. PDMS assigns zero progress outside $A = \{\mathrm{NC} \cdot \mathrm{DAC} > 0\}$, so

$$\mathbb{E}[\mathrm{EP}_{\mathrm{rep}}] = \Pr(A)\,\mathbb{E}[\mathrm{EP}_{\mathrm{rep}} \mid A]. \tag{12}$$

For NoRD, the pass rate rises by about 11 percentage points while conditional progress remains nearly unchanged, increasing reported progress despite slower, shorter plans. Qwen-Drive improves both the pass rate and conditional progress. Interpreting a progress gain therefore requires separating these two contributions.

Broader coverage and paired-state evaluation. Expanding what is scored exposes additional trade-offs but does not ensure that every behavior improves with the aggregate. When fixed outputs are rescored with EPDMS, both ReCogDrive pairs retain score gains while lane keeping and history comfort decrease; DiffusionDrive v2's score advantage becomes much smaller, with an interval containing zero, while history comfort also decreases (Table 1). TOAD's deployed search changes PDMS by $+0.001$ and EPDMS by -0.003, with the PDMS log interval including zero; NoRD improves all displayed EPDMS components. The separate official two-stage evaluation generates predictions at original and synthetic states and measures extended comfort from time-aligned simulated motion in consecutive plans. ReCogDrive and Qwen-Drive improve drivable-area compliance and progress in both stages while extended comfort falls, but the combined-score changes of all three have paired intervals containing zero; NoRD improves all displayed components and retains a positive combined gain (Table 3). Broader coverage reveals these inconsistencies, while the

preceding motion-layer tests leave a further question: whether gains survive a change in execution during repeated replanning.

4 RL Gains Depend on the Execution Interface

The benefit of optimization depends on the process that executes requests and returns them as feedback. Section 3.4 shows that better tracked scores can coexist with poorer requested-motion diagnostics and weak responses to substantial trajectory edits. The closed-loop recurrence in Eq. 6 adds another consequence: execution changes the inputs on which later requests depend. We test whether gains persist across execution interfaces, then examine how successive predictions change during replanning. We compare LQR with a bicycle model against playback on matched WorldEngine scenarios, keeping non-reactive traffic, scoring, rendering settings, and inference seeds fixed. Initial predictions agree exactly across interfaces for each checkpoint. This isolates the effect of execution and its subsequent feedback; protocol details and behavioral checks appear in Appendix D.1.

Table 4: **Gain reversals across execution interfaces.** Mean PDMS on matched NR scenes per family. Ref./Opt.: IL/base→RL, except TOAD Frozen→CEM; Δ: Opt.−Ref. Interaction: $I_{\mathrm{LQR,PB}} = \Delta S_{\mathrm{LQR}} - \Delta S_{\mathrm{PB}}$ (Eq. 7).

	LQR + bicycle model			Playback			LQR−PB		Interaction
Planner	Ref.	Opt.	Δ	Ref.	Opt.	Δ	Opt.	Ref.	$I_{\mathrm{LQR,PB}}$ [95% CI]
ReCogDrive-2B	0.580	0.798	**+0.218**	0.688	0.452	**−0.236**	**+0.346**	**−0.108**	**+0.454** [+0.377, +0.533]
ReCogDrive-8B	0.628	0.786	**+0.158**	0.690	0.490	**−0.200**	**+0.295**	**−0.062**	**+0.358** [+0.274, +0.436]
Qwen-Drive	0.697	0.834	**+0.136**	0.787	0.398	**−0.388**	**+0.435**	**−0.089**	**+0.525** [+0.448, +0.603]
NoRD	0.462	0.648	**+0.186**	0.516	0.672	**+0.156**	**−0.024**	**−0.054**	+0.030 [−0.044, +0.102]
TOAD	0.843	0.916	**+0.073**	0.761	0.854	**+0.093**	**+0.063**	**+0.082**	−0.020 [−0.062, +0.022]

For both ReCogDrive sizes and Qwen-Drive, switching from playback to LQR lowers the imitation baseline's score but raises the RL checkpoint's score (Table 4, LQR−PB columns). Changing execution reverses the sign of the RL gain for both ReCogDrive sizes and Qwen-Drive (Table 4). All three improve under LQR tracking and deteriorate under playback, with log-bootstrap intervals excluding zero in both directions; NoRD and TOAD retain positive gains under both. The interaction intervals exclude zero for all three reversing pairs; Qwen-Drive's change in RL gain is $+0.525$. The benefit of optimization therefore depends on the execution interface. For Qwen-Drive, road-departure counts rise from 26 to 111 after RL under playback, but fall from 66 to 27 under LQR. Under playback, its IL-to-RL drivable-area compliance falls from 0.91 to 0.61 and ego progress from 0.68 to 0.40 (Table 11).

A single-plan control distinguishes executing a fixed request from repeated replanning. In a separate Qwen-Drive control with the same first plan frozen, replacing LQR with playback changes the RL score by only -0.005, with a log-bootstrap 95% interval of $[-0.035, +0.024]$. This small contrast motivates examining subsequent requests and returned states; the controls are not an additive causal decomposition (Appendix D.1).

Replanning can repeatedly replace a predicted slowdown with a higher speed request. To distinguish these effects, let $\hat{v}_k(r)$ denote longitudinal speed at relative time r in plan k, and let $h = 0.5\,\mathrm{s}$ be the replanning interval. Under playback, the input speed is $v_k = \hat{v}_{k-1}(h)$, giving, for $k \geq 1$,

$$v_{k+1} - v_k = \underbrace{\hat{v}_{k-1}(2h) - \hat{v}_{k-1}(h)}_{\text{previous plan's speed change}} + \underbrace{\hat{v}_k(h) - \hat{v}_{k-1}(2h)}_{\text{same-time replanning revision}} . \tag{13}$$

The second term compares consecutive predictions for the same absolute future time, separating a change already intended by the previous plan from a revision introduced by replanning.

The earlier matched Qwen-Drive NR playback rollouts show a persistent contrast between these two quantities. Of 2,296 RL plans, 1,897 (82.6%) predict slowing down from h to $2h$, yet all seven same-time revisions are positive in 196 RL scenes, compared with one IL scene. Mean revisions are $+2.184\,\mathrm{m/s}$ for RL and $-0.262\,\mathrm{m/s}$ for IL; the saved records satisfy Eq. 13 to numerical precision (Appendix D.3). They show how slowing within individual plans can coexist with repeated upward revisions during execution.



Figure 3: **Qwen-Drive at the same junction.** Historical WorldEngine non-reactive playback front-camera renders with timestamps and longitudinal speeds.

Successive predictions show that the execution interface also changes the planner's response to feedback. On the full interface cohort, the same-time revision $\hat{v}_k(h) - \hat{v}_{k-1}(2h)$ averages $+2.17$ m/s for RL under playback, versus $+0.58$ under LQR; the IL changes are -0.27 and -0.06 m/s, respectively. The paired change in RL's speed revision is -1.58 m/s, with a log-bootstrap 95% interval of $[-1.71, -1.45]$. The matched intervention therefore changes subsequent requests as well as executed motion; the speed revisions describe this response without isolating its causal contribution to the score reversal. In the earlier playback evaluation, Qwen-Drive's RL planner repeatedly replaces a predicted slowdown with a higher near-term speed and leaves the road (Figure 3). The IL planner initially stops, then continues toward the same junction while remaining on road beyond the evaluation horizon. ReCogDrive's earlier playback records likewise show more frequent predicted forward–backward motion after RL (Table 13). These examples illustrate successive requests; the matched comparison establishes the gain reversal.

Interpretation. The matched results establish that RL gains can depend on the execution interface and its subsequent feedback. One possible explanation is that tracking suppresses request variations whose consequences emerge under playback, but the contribution of specific filtered components remains to be isolated (Appendix D.1). An improvement claim therefore concerns both the planner's requests and the process that executes, scores, and returns them as feedback.

5 Conclusions & Take-aways

Our results show that metric validity under optimization cannot be inferred from aggregate score gains alone. Behavioral distinctions can be lost through omission, thresholding, and transformations between requested and executed motion. Gains can also reverse across execution interfaces during repeated replanning, showing that the benefit of optimization depends on how requests are executed and returned as feedback. These findings motivate evaluating both the behavioral distinctions preserved by the score and the execution and feedback process in which optimized outputs are used.

Our take-home messages are:

1. **Analyze the scoring chain before making it a target.** Identify which behaviors are omitted, which differences receive the same score, and how execution changes the motion being assessed. Test these potential blind spots with controlled interventions.
2. **Preserve the behavioral quantities behind the aggregate.** Report continuous measurements alongside thresholded scores, distinguish requested from executed motion, and report progress together with the safety constraints that govern its score.
3. **Evaluate gain transfer under execution and feedback.** Report how planner outputs are executed. Test whether gains persist when execution changes and planning repeatedly responds to the resulting states, while tracking safety-critical failures.
4. **Revalidate the complete scoring process.** Changing scoring rules does not by itself address what execution alters or removes. Check whether the revised process captures meaningful behavioral differences and whether optimizing it produces better behavior.

AI USE STATEMENT

We used large language models (ChatGPT/Claude) to improve the English phrasing and readability of the manuscript and assist with experimental code development and execution. The research ideas, proofs, and core scientific arguments were developed by the authors. The authors take full responsibility for the final content of this paper.

REFERENCES

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. In *CVPR ADP3 Workshop*, 2021. arXiv:2106.11810.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *CoRL*, 2025. arXiv:2506.04218.

Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *CoRL*, 2023. arXiv:2306.07962.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *NeurIPS Datasets and Benchmarks*, 2024. arXiv:2406.15349.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *NeurIPS Datasets and Benchmarks*, 2024. arXiv:2406.03877.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning. In *ICLR*, 2024. arXiv:2310.09144.

Tianyu Li, Li Chen, Caojun Wang, Haochen Liu, Kashyap Chitta, Zhenjie Yang, Yuhang Lu, Naisheng Ye, Yihang Qiu, Yufei Wang, Luoxi Zou, Jiaxin Peng, Jin Pan, Zhaoyu Su, Andrei Bursuc, Shengbo Eben Li, Andreas Geiger, Peng Su, and Hongyang Li. World Engine: Towards the era of post-training for autonomous driving, 2026a. URL https://arxiv.org/abs/2606.19836. Technical report.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. ReCogDrive: A reinforced cognitive framework for end-to-end autonomous driving. In *ICLR*, 2026b. arXiv:2506.08052.

Zhenxin Li, Wenhao Yao, Zi Wang, Xinglong Sun, Joshua Chen, Nadine Chang, Maying Shen, Zuxuan Wu, Shiyi Lan, and José M. Álvarez. Generalized trajectory scoring for end-to-end multimodal planning. *arXiv:2506.06664*, 2025.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. DiffusionDrive: Truncated diffusion model for end-to-end autonomous driving. In *CVPR*, 2025. arXiv:2411.15139.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv:1803.04585*, 2018.

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.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Xiaolong Tang, Meina Kan, Shiguang Shan, and Xilin Chen. Plan-R1: Safe and feasible trajectory planning as language modeling. In *ICLR*, 2026. arXiv:2505.17659.

Yiru Wang, Anqing Jiang, Shuo Wang, Yuwen Heng, Hai Yang, Yang Chen, and Hao Sun. Do open-loop metrics predict closed-loop driving? A cross-benchmark correlation study of NAVSIM and Bench2Drive. *arXiv:2605.00066*, 2026.

Yihong Xu, Éloi Zablocki, Yuan Yin, Elias Ramzi, Ellington Kirby, Alexandre Boulch, and Matthieu Cord. Test-time trajectory optimization for autonomous driving. *arXiv:2606.07170*, 2026.

Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuScenes. *arXiv:2305.10430*, 2023.

Yinan Zheng, Ruiming Liang, Kexin Zheng, Jinliang Zheng, Liyuan Mao, Jianxiong Li, Weihao Gu, Rui Ai, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Diffusion-based planning for autonomous driving with flexible guidance. In *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5c0cd43063f087eb32e572d040a19cd2-Abstract-Conference.html.

Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. HUGSIM: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *arXiv:2412.01718*, 2024.

Xin Zhou, Zongchuang Zhao, Zhibo Yang, Mingsheng Li, Humen Zhong, Shuai Bai, Du Chu, Ruizhe Chen, Zhaohai Li, Jun Tang, Qiuyue Wang, Mingkun Yang, Jiazhao Zhang, Dayiheng Liu, Dingkang Liang, and Xiang Bai. Qwen-Drive-1.0: An initial step towards a vision-language foundation model for autonomous driving. *arXiv preprint arXiv:2609.00111*, 2026.

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:2512.07745*, 2025.

A Evaluation Protocols and Diagnostics

A.1 Scene Sets and Comparisons

We use the scene sets defined in Section 3.1; the WorldEngine test set follows the official split (Li et al., 2026a). We use the released `navtest_failures_filtered.yaml` scene list at WorldEngine revision `95002ad`, which contains failure-prone scenarios from `navtest`. Open-loop scoring evaluates a four-second plan after LQR tracking with a kinematic bicycle model; surrounding agents replay their logs. WorldEngine simulation provides observations from the evolving ego state over a four-second rollout with eight updates at 0.5 s intervals, following four logged history frames; we score the resulting rollout post hoc. The released entry point uses trajectory playback. The client converts rear-axle poses to world-frame vehicle-center poses and obtains velocities by forward differences on the 10 Hz trajectory. At each update it applies the pose and motion state at $h = 0.5$ s and returns the resulting state and rendered observations for replanning. In particular, the returned velocity is $[p_c(h + 0.1) - p_c(h)]/0.1$, where p_c is the vehicle-center position. ReCogDrive's 2 Hz plans are linearly interpolated to this 10 Hz grid; Qwen-Drive emits native 10 Hz poses. Both Qwen-Drive checkpoints use reasoning-conditioned inference with one sampled trajectory and seed 42. The evaluation records closed-loop PDMS, at-fault collisions, and drivable-area exits. The progress guard is 5.0 m in open loop and 0.1 m in closed loop, so absolute EP values are not compared across regimes. For v1 EP normalization, route progress is first multiplied by the multiplicative penalties. If the maximum of these adjusted progress values across the evaluated proposal and PDM reference exceeds the guard, each is divided by that maximum. Otherwise, EP is one for proposals with a nonzero multiplicative penalty product and zero for those with a zero product.

Open-loop pairs use shared scenes. The primary execution-interface experiment uses 288 scenes in every condition (Appendix D.1). Earlier playback comparisons use 286 jointly completed ReCogDrive scenes, one fewer than the full set for Qwen-Drive, and the full set for NoRD/TOAD; these historical cohorts remain separate. The nuPlan replication uses `val14` in both traffic regimes (Section 3.1).

A.2 Behavioral Measurements

Sampling and geometry. NAVSIM diagnostics use eight predicted positions $p_1, \ldots, p_N$ at $\Delta t = 0.5\,\text{s}$, with $N = 8$. Smoothing displacement follows Equation 10, applying one three-point moving average with the first and last predicted positions held fixed. Its unit is meters; smaller values mean a smaller required positional edit. The local difference equals one third of the second positional difference and can reflect both bending and uneven speed. For heading jitter and path length, prepend the ego origin p_0 and compute

$$\theta_i = \operatorname{unwrap}\big(\operatorname{atan2}(y_i - y_{i-1}, x_i - x_{i-1})\big), \qquad L = \sum_{i=1}^{N} \|p_i - p_{i-1}\|_2. \tag{14}$$

Heading jitter is the sum in Equation 10, in radians; it is not a time derivative. Path length contextualizes headings when displacements are small.

Longitudinal and lateral decomposition. The supplementary Qwen decomposition prepends p_0 before smoothing, so its first internal point has a different endpoint convention from D_{smooth}. Let δp_i be the edit on this extended sequence and t_i the unit direction of the incoming segment, with perpendicular n_i. The reported components are

$$D_{\parallel} = \max_i |\delta p_i^\top t_i|, \qquad D_{\perp} = \max_i |\delta p_i^\top n_i|. \tag{15}$$

The implementation divides each segment by its length plus 10^{-9} to handle zero displacement. These maxima are separate directional readouts, not additive components of the main-table maximum.

Pace and acceleration. Planned segment speed and speed change are

$$v_i = \frac{\|p_i - p_{i-1}\|_2}{\Delta t}, \qquad a_i = \frac{v_{i+1} - v_i}{\Delta t}. \tag{16}$$

The Qwen pace table reports signed extrema $\max_i a_i$ and $-\min_i a_i$ as acceleration and deceleration, without clipping them at zero. ReCogDrive's plan/execution comparison and the retiming experiment use the absolute planned peak $\max_i |a_i|$. The supplementary planned lateral quantity is $\max_i |y_{i+1} - 2y_i + y_{i-1}|/\Delta t^2$ in the initial ego frame. Longitudinal displacement is $(p(t) - p_0)^\top (\cos\psi_0, \sin\psi_0)$; its time derivative describes longitudinal pace in this fixed frame. Executed-motion quantities are computed from simulated states separately.

Plan adherence and replanning. NAVSIM interpolates the requested trajectory to the simulator's 0.1 s grid and reports Equation 11 over the four-second horizon, including the initial state. For nuPlan, let $p^{(t)}(u)$ be the position requested at time t for absolute time u. At each valid planning step t_i, compare the plan with the later executed positions at the available horizons $h \in \{0.5, 1, \ldots, 4\}$ s:

$$e_{i,h} = \|\hat{p}(t_i + h) - p^{(t_i)}(t_i + h)\|_2, \quad D_{\text{adh,max}} = \max_{i,h} e_{i,h}, \quad D_{\text{adh,mean}} = \frac{1}{M} \sum_i \max_h e_{i,h}. \tag{17}$$

Later nuPlan execution incorporates replanning, so this measurement includes departures induced by subsequent plans. For plans issued 0.5 s apart, churn compares matching absolute times:

$$C_i = \frac{1}{7} \sum_{k=1}^{7} \|p^{(t_i)}(t_i + 0.5 + 0.5k) - p^{(t_i+0.5)}(t_i + 0.5 + 0.5k)\|_2. \tag{18}$$

The scene value is the mean of C_i over valid pairs, with the maximum retained as a supplementary quantity. Missing plans are excluded, and quantities requiring a planned trajectory are inapplicable to action interfaces without one. NuPlan plan-geometry scene values average over valid replanning steps.

Executed motion and comfort. For an official acceleration or jerk time series $z_r(t)$, an absolute peak is $\max_t |z_r(t)| = \max\{|\min_t z_r(t)|, |\max_t z_r(t)|\}$; this is the operator used for the ReCogDrive comparison and nuPlan jerk results. Qwen's simulated acceleration and deceleration columns

instead report $\max_t a_\parallel(t)$ and $-\min_t a_\parallel(t)$, matching its signed planned extrema. Equation 9 defines longitudinal jerk; numerical filtering and differentiation follow the official metric implementation. Comfort checks six channels: longitudinal and lateral acceleration, jerk magnitude, longitudinal jerk, yaw acceleration, and yaw rate. With channel bounds ℓ_r, u_r, the NAVSIM binary test is

$$C(\tau) = \prod_{r=1}^{6} \mathbf{1}[\ell_r < z_r(t) < u_r \text{ for all evaluated } t]. \tag{19}$$

Continuous replacements use the bound-utilization ratios r_k specified in Appendix B.2. History comfort applies the same test to the concatenation of logged past states and simulated future states, $\mathrm{HC} = C(\tau_{\mathrm{past}} \oplus \tau_{\mathrm{exec}})$. For two successive simulated predictions aligned at their overlapping absolute times, let $z_r^{(1)}(t_i)$ and $z_r^{(2)}(t_i)$ be acceleration magnitude, jerk magnitude, yaw rate, or yaw acceleration. The official extended-comfort term is

$$\mathrm{EC} = \prod_r \mathbf{1}\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z_r^{(1)}(t_i) - z_r^{(2)}(t_i)\right)^2} \le b_r\right]. \tag{20}$$

The four bounds are $0.7\,\mathrm{m/s^2}$, $0.5\,\mathrm{m/s^3}$, $0.1\,\mathrm{rad/s}$, and $0.1\,\mathrm{rad/s^2}$, respectively. The two-stage experiment uses this term for the aligned prediction pairs in the official protocol (Appendix E.2).

Driving direction. Let I_i^{opp} be the scorer's oncoming-traffic mask. The largest accumulated displacement in its one-second sliding window is

$$d_{\mathrm{opp}} = \max_j \sum_{i=\max(0,j-H)}^{j} \|p_i - p_{i-1}\|_2 I_i^{\mathrm{opp}}, \tag{21}$$

where the initial increment is zero and $H = 1\,\mathrm{s}/\Delta t_{\mathrm{sim}}$. The score is 1 for $d_{\mathrm{opp}} < 2\,\mathrm{m}$, 0.5 for $2 \le d_{\mathrm{opp}} < 6\,\mathrm{m}$, and 0 otherwise. The v2 implementation removes intersection positions from the mask; the v1 and v2 DDC values are therefore reported separately. Map-verified wrong-way distance in the controlled probe is a separate measurement from this thresholded score.

Lane keeping. Let $d_i = \mathrm{dist}(p_i, \mathcal{C})$ for the route centerline $\mathcal{C}$. Starting from $q_0 = 0$, the implementation updates the exceedance counter as

$$q_i = \begin{cases} q_{i-1}, & p_i \text{ is in an intersection,} \\ q_{i-1} + 1, & d_i > 0.5\,\mathrm{m} \text{ outside intersections,} \\ 0, & \text{otherwise,} \end{cases} \qquad \mathrm{LK} = \mathbf{1}\left[\max_i q_i < \left\lceil \frac{2\,\mathrm{s}}{\Delta t_{\mathrm{sim}}} \right\rceil\right]. \tag{22}$$

This differs from drivable-area compliance, which checks the vehicle footprint against the drivable region.

Outcome rates and reported progress. For outcome f (at-fault collision or road departure), let I_s^f indicate its occurrence in scene s. Over n matched scenes, the count is $N_f = \sum_{s=1}^{n} I_s^f$ and the rate is $r_f = N_f/n$. They are distinct from mean penalty values, since NC includes a partial penalty for static-object collisions. Equation 12 defines the reported-progress decomposition; comparisons conditional on both arms passing use their intersection, rather than each arm's own passing set. All paired differences are optimized minus baseline in the metric's native units.

A.3 Descriptive Planner Profiles

Figure 1(b) groups 24 configurations into IL (13), RL (5), score distillation (5), and CEM (1): 22 checkpoints and the TOAD checkpoint evaluated without and with search. NoRD-SFT belongs to IL; the Qwen-Drive pair is reported separately in Table 1. Configurations using multiple procedures are assigned once with priority CEM, RL, distillation, then IL. Each profile averages configuration-level ratios with equal weight; planner composition differs across groups, so the profiles are descriptive.

For scored components, lane keeping, and direction, q is the planner mean divided by the logged-human mean on the same scenes. For smoothing displacement and plan-adherence error, the ratio is

reversed. Human $= 1$ denotes equality with the logged-human mean on the same scenes. All axes map ratios $(0, 0.5, 0.8, 1.1, 1.27)$ to radii $(0, 0.12, 0.28, 0.88, 1)$ by piecewise-linear interpolation. The upper half contains components that influence PDMS; the lower half contains zero-weight DDC, single-stage NAVSIM-v2 lane keeping, smoothing displacement, and plan adherence.

A.4 Paired Statistical Analysis

Mean changes use paired scene-bootstrap 95% intervals with 2,000 resamples. Log-bootstrap intervals resample driving logs, retaining their scenes together to account for within-log dependence. Wilcoxon signed-rank tests use Holm correction over the original analysis families: 42 geometry/score tests and 30 EPDMS tests for the six training comparisons, and 24 tests across TOAD's augmentation and search arms. The displayed tables select quantities relevant to the main text; these selections do not change the correction families. A small signed-rank p-value need not imply a nonzero mean change, so confidence intervals guide interpretation of mean effects. Qwen-Drive's DDC change, NoRD's plan-adherence change, GTRS-Dense's jitter change, and TOAD's `navtest` PDMS change have log intervals including zero. We retain these cases alongside supported improvements and declines.

In Table 1, † marks changes for which either interval includes zero, the interval directions disagree, or the Holm-adjusted test does not pass $p < 0.05$. For nuPlan, support likewise requires the scene and log intervals to exclude zero in the same direction and the Wilcoxon test to pass Holm correction over the original 38-quantity family in each traffic regime. The family is retained even for measurements omitted from the final tables. These intervals quantify evaluation-scene and driving-log variability for the fixed released checkpoints.

A.5 Planner-Specific Settings

NoRD. NoRD uses a fixed human-derived trajectory vocabulary and PDMS rewards during reinforcement fine-tuning (Rawal et al., 2026). This specifies the action representation for the comparison; the observed diagnostic changes are reported in Table 1.

Qwen-Drive. The reinforcement objective uses EP/TTC/comfort weights $6/4/2$ and a displacement reward toward the logged trajectory, with exploration in six low-frequency cosine modes (Zhou et al., 2026). Both released experts use reasoning-conditioned inference, matching the reinforcement checkpoint's training mode; direct SFT inference supplies the conditioning control in Appendix B.4. The expert emits 50 poses at 10 Hz, from which eight poses at 0.5 s intervals are selected for scoring. Inputs use the three released camera views at the four most recent 2 Hz frames and 10 Hz ego history from the nuPlan database, filtered according to the released model documentation. Scoring the leading 40 poses instead changes PDMS by 0.03 points. For WorldEngine replay, both experts use the same four rendered camera frames, with the required 10 Hz ego history interpolated from the available 2 Hz states; the leading 40 predicted poses are passed to the standard trajectory interface. Repeated deterministic inference reproduces the trajectories bit-identically. Reproduced PDMS is 87.7, 87.9, and 90.5 for direct SFT, reasoning-conditioned SFT, and RL, respectively, against published values of 87.8, 88.2, and 90.7; all sub-metrics agree within 0.5 points (Zhou et al., 2026).

TOAD on NAVSIM-v1. The released DrivoR checkpoint uses 64 planner proposals, 64 random control-space perturbations, and CEM with 64 samples, 8 elites, and 10 rounds. Perturbations are divided equally among constant, linearly varying, and free-form controls. Search combines predicted sub-scores through the PDMS rule, with zero DDC weight, a comfort penalty of weight 0.05, and an anchoring penalty. The CEM mean is deployed only when its predicted score exceeds that of the best proposal. We seed perturbations per scene from its token, making the baseline, augmentation-only, and search arms independent of processing order. All recorded trajectories are evaluated by the same official scorers. The v2 direction comparison and search configuration are specified in Appendix C.3.

Plan-R1 on nuPlan. Evaluation uses 15 s scenarios, 10 Hz replanning, LQR execution, and logged or IDM-reactive traffic. The first plan is also evaluated without replanning using the CLS-style PDM score. Both checkpoints share a 1,024-token vocabulary and a frozen prediction branch; all 482

tensors of that branch match between the released files. Fine-tuning samples $G = 4$ sequences and uses

$$r_t = g_t^{\text{road}} g_t^{\text{agent}} g_t^{\text{obstacle}} \frac{2c_t + 5\tau_t + 2s + p}{10}, \tag{23}$$

where the gates penalize road departures and collisions, c_t and τ_t are comfort and TTC, and s and p are rollout speed-limit compliance and progress. Group-centered rewards-to-go use fixed scaling 0.1; the KL penalty toward the frozen branch also has weight 0.1. The reward changes CLS weights and omits its making-progress and direction gates; comfort remains thresholded, without a continuous jerk penalty. Executed peaks use $\max_t |x_t|$, computed from both extrema of each time series. Plan geometry uses each plan's first four seconds; churn compares plans issued 0.5 s apart at matching absolute times.

B Metric Sensitivity

B.1 Comfort and Tracker Effects

Binary comfort fails on 410 of 291,504 `navtest` planner–scene evaluations (0.14%). One of its six checks is constant because the motion model's lateral-acceleration channel is zero. For DiffusionDrive, DiffusionDriveV2, and the human trajectory on `navhard`, the largest within-scene comfort utilization averages 0.32–0.38 of its threshold. Figure 4 relates this margin to the difference between planned and tracked trajectories. For ReCogDrive-2B, the mean absolute planned acceleration peak rises from 1.14 to 2.84 m/s^2 after fine-tuning, while the absolute executed longitudinal peak rises from 0.88 to 1.20 m/s^2. The corresponding 8B values are 1.14 to 2.95 planned and 0.89 to 1.34 executed.

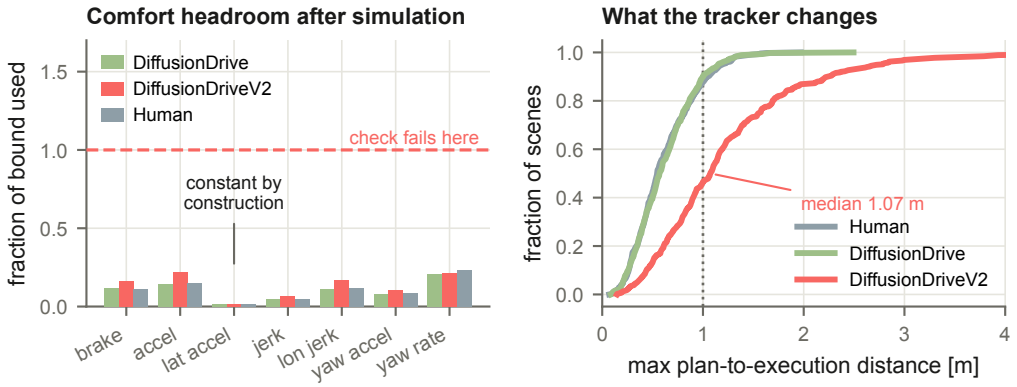

Figure 4: **Comfort sensitivity and plan adherence on `navhard`. Left:** Mean utilization of each comfort bound for DiffusionDrive v1, v2, and logged Human trajectories; a value of one denotes the bound. The lateral-acceleration check is constant by construction. **Right:** Empirical cumulative distribution of the maximum frame-wise distance between planned waypoints and LQR-executed positions, in meters.

B.2 Continuous Comfort Rescoring

For continuous rescoring, the two-sided longitudinal-acceleration check is separated into braking and propulsion, giving seven bound utilizations r_k. For an upper bound $b_k > 0$, $r_k = \max(0, \max_t z_k(t)/b_k)$; for the lower braking bound $-b_k$, $r_k = \max(0, -\min_t a_\parallel(t)/b_k)$; magnitude checks use $r_k = \max_t |z_k(t)|/b_k$. The bounds for braking, propulsion, lateral acceleration, jerk magnitude, longitudinal jerk, yaw acceleration, and yaw rate are 4.05, 2.40, 4.89, 8.37, 4.13, 1.93, and 0.95, in their corresponding SI units. Using the same fixed trajectories, we replace binary

comfort by three functions of these utilizations:

$$C_{\text{worst}} = \text{clip}(1 - \max_k r_k, 0, 1), \tag{24}$$

$$C_{\text{sigmoid}} = \sigma\big((1 - \max_k r_k)/0.1\big), \tag{25}$$

$$C_{\text{mean}} = K^{-1} \sum_k \text{clip}(1 - r_k, 0, 1). \tag{26}$$

Here $K = 7$ and $\sigma(x) = 1/(1 + e^{-x})$; all other PDMS components are unchanged. The ranking of DiffusionDrive, DiffusionDriveV2, and Human remains unchanged (Table 5); the worst-channel margin retains 65% and 75% of the original DiffusionDriveV2 advantage on the two sets. This fixed-output comparison measures how each comfort definition separates the three planners.

Table 5: **Comfort rescoring on `navhard` and the WorldEngine test set.** Mean PDMS (↑) under the official and continuous comfort formulations, using fixed outputs. DD denotes DiffusionDrive and DDv2 denotes DiffusionDriveV2. Retained reports the DDv2–DD difference relative to the official binary difference, computed from unrounded means.

	`navhard`				WorldEngine test set			
Comfort term	DD	DDv2	Human	Retained	DD	DDv2	Human	Retained
Official (binary)	0.7044	0.7328	0.8992	100%	0.6540	0.6960	0.9325	100%
Worst-channel margin	0.6616	0.6801	0.8458	65%	0.6144	0.6458	0.8778	75%
Smooth sigmoid	0.7038	0.7315	0.8982	97%	0.6534	0.6948	0.9313	98%
Mean margin	0.6903	0.7138	0.8814	83%	0.6410	0.6777	0.9143	87%

B.3 Driving-Direction Construction

A route-centerline follower is compared with offsets of $\{\pm 1.75, \pm 3.5\}$ m at the same speed profile. Matched trajectories pass NC and DAC, differ in raw progress by less than 0.1 m, and have identical TTC and comfort. Lane direction at flagged frames is verified against map lane baselines, accounting for the different traffic conventions in Singapore and the US maps. The construction yields 42 pairs across 37 `navhard` scenes, with 8.80 m of mean verified wrong-way travel; DDC is zero for 31 of the 42 violating trajectories. Mean paired PDMS differs by $+0.000019$ and the median difference is zero. Re-aggregating the same sub-scores with DDC weight 2 or 5 gives differences of -0.157 or -0.278; using DDC as a gate gives -0.797.

B.4 Positional Smoothing and Re-timing

Positional smoothing applies one three-point moving average to waypoints with fixed endpoints and preserves their original headings. Holding headings fixed isolates the positional edit from heading reconstruction. For DiffusionDriveV2 it removes 70% of heading jitter on `navhard` and 74% on the WorldEngine test set, with mean waypoint displacements of 0.417 and 0.409 m. PDMS changes by -0.003 $[-0.013, +0.007]$ and $+0.007$ $[-0.006, +0.021]$, respectively. The same edit changes DiffusionDrive PDMS by $+0.004$ and -0.002, and Human PDMS by -0.005 and -0.016 on the respective sets.

Qwen-Drive controls. Switching SFT from direct to reasoning-conditioned inference changes PDMS by $+0.002$ $[+0.001, +0.004]$, smoothing displacement by $+0.000$ m, and plan-adherence error by -0.001 m. This conditioning comparison does not account for the smoothing-displacement increase between the SFT and RL checkpoints in Table 6. The larger longitudinal component locates most of the variation along the path.

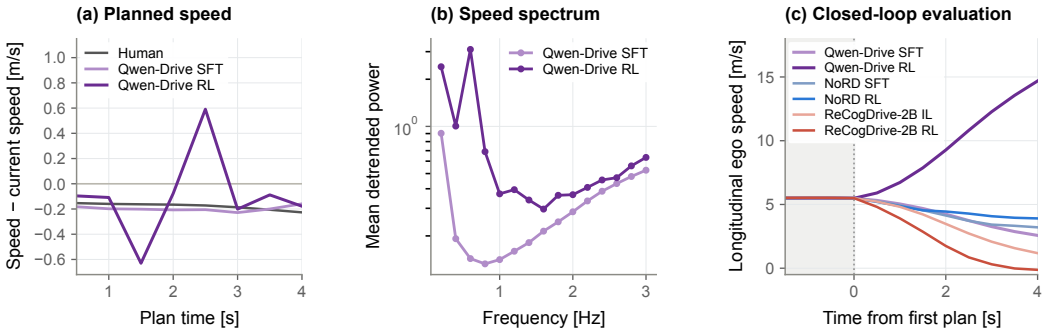

Figure 5: **Planned and executed speed.** (a) Mean planned speed relative to ego speed on `navtest`. (b) Mean speed spectrum after linear detrending (5 s, 10 Hz). (c) Mean longitudinal speed in WorldEngine NR simulation (Li et al., 2026a); matched sets: Appendix A. Shading: logged history; zero: first action.

Table 6: **NAVSIM-v1 `navtest`: plan and simulated motion.** Geometry and pace of the Qwen-Drive experts and the human trajectory: smoothing displacement and its components along and across the local path direction, planned peak acceleration and deceleration between consecutive 0.5 s segments, and the peak acceleration and deceleration of the simulated states that the scorer evaluates.

	Smoothing displacement [m] ↓			Planned [m/s^2]		Simulated [m/s^2]	
Checkpoint	Total	Along	Across	Acc.	Dec.	Acc.	Dec.
Qwen-Drive SFT	0.116	0.103	0.049	0.67	0.66	0.32	0.41
Qwen-Drive RL	0.212	0.204	0.059	1.65	1.88	0.28	0.53
Human	0.099	0.081	0.050	0.37	0.43	0.32	0.43

Re-timing smooths segment speeds along each trajectory's own path, preserving endpoint speeds and total path length. On `navtest` it reduces the RL checkpoint's largest planned speed change per unit time from 2.46 to 1.31 m/s^2, with mean PDMS change -0.0002 and an interval including zero (Table 7). The SFT and Human controls also have small mean score changes under this intervention.

Table 7: **NAVSIM-v1 `navtest`: retiming intervention.** Re-timing the deployed trajectory along its own path: largest speed change per unit time between consecutive segments before and after, PDMS before and after, and the paired change with its bootstrap 95% interval.

	$\max \lvert\Delta v\rvert/\Delta t$ [m/s^2] ↓		PDMS ↑		
Checkpoint	Original	Retimed	Original	Retimed	Δ [95% CI]
Qwen-Drive SFT	1.24	0.91	0.8791	0.8790	−.0002 [−.0005, .0002]
Qwen-Drive RL	2.46	1.31	0.9047	0.9045	−.0002 [−.0008, .0003]
Human	0.97	0.87	0.9455	0.9453	−.0002 [−.0005, .0000]

B.5 Progress and Candidate Ties

Equation 12 separates reported EP into pass frequency and conditional progress. NoRD's EP increases by 0.091 as its pass rate rises from 81.7% to 92.3%, while conditional EP changes from 0.846 to 0.847. Its mean planned speed falls by 0.16 m/s and path length by 0.66 m. Qwen-Drive instead gains 0.027 EP with both pass rate (94.9% to 96.8%) and conditional EP (0.865 to 0.876) increasing; mean speed rises by 0.10 m/s and path length by 0.40 m. These comparisons distinguish a change in pass frequency from a change in progress among passed scenes.

In the fixed 800-candidate DiffusionDriveV2 pool on `navhard`, 274 scenes have tied PDMS maxima and 269 have tied EPDMS maxima. Mean tied-set sizes are 146.6 and approximately 137. Within the respective tied sets, heading jitter spans 0.43–7.29 and 0.45–5.63, while peak lateral acceleration spans 0.98–6.29 and 0.99–5.61 m/s^2. Both scores therefore leave geometric differences unresolved among some maximizers.

C Paired Optimization Responses

C.1 Scene-Set Replications

Table 8 combines the `navhard` and WorldEngine test-set replications, reporting PDMS changes alongside fixed-output EPDMS rescoring and behavioral diagnostics. ReCogDrive's geometric changes and NoRD's reductions in jitter and smoothing displacement recur on both sets. Qwen-Drive again increases smoothing displacement while reducing plan-adherence error; the GTRS-Dense changes remain small on these quantities.

Table 8: **Paired changes on safety-critical cases.** Scores, EPDMS components, and behavioral diagnostics on fixed planner outputs from `navhard` and the WorldEngine test set. PDMS uses the v1 scorer; EPDMS and its components use the v2 scorer. DDC^{v1} has zero PDMS weight, whereas DDC^{v2} gates EPDMS and excludes intersection positions. Jitter is in radians; smoothing displacement (Smooth.) and plan-adherence error (Adh.) are in meters. Gray entries have unadjusted paired-test $p \geq 0.05$ in these subset analyses; `navtest` adjusted tests and intervals appear in Tables 9 and 18.

	Aggregate scores		EPDMS components			Behavioral diagnostics			
Pair	PDMS ↑	EPDMS ↑	DDC^{v2} ↑	LK ↑	HC ↑	DDC^{v1} ↑	Jitter ↓	Smooth. ↓	Adh. ↓
navhard									
ReCog-2B IL→RL	+.105	+.077	−.029	−.044	−.004	−.032	+0.83	+.154	+.385
ReCog-8B IL→RL	+.079	+.060	−.016	−.024	−.004	−.012	+0.71	+.164	+.399
Qwen-Drive SFT→RL	+.063	+.062	−.001	+.016	+.000	−.001	+.03	+.091	−.067
DDv1→v2	+.028	−.000	−.017	−.011	−.096	−.024	+1.25	+.310	+.609
NoRD IL→RL	+.152	+.145	+.050	+.056	+.031	+.046	−0.29	−.052	−.053
GTRS-Dense expert→reward	+.004	−.003	−.011	−.022	+.007	−.014	−.05	−.001	−.089
WorldEngine test set									
ReCog-2B IL→RL	+.093	+.079	−.028	−.028	.000	−.035	+1.07	+.156	+.358
ReCog-8B IL→RL	+.057	+.034	−.009	−.014	−.003	−.010	+0.84	+.167	+.412
Qwen-Drive SFT→RL	+.050	+.043	+.000	+.007	+.003	−.003	+.07	+.092	−.066
DDv1→v2	+.042	+.014	−.028	−.003	−.094	−.030	+1.46	+.300	+.581
NoRD IL→RL	+.159	+.139	+.061	+.087	+.045	+.045	−0.34	−.064	−.074
GTRS-Dense expert→reward	+.013	−.007	−.009	−.024	+.014	−.021	+.05	−.004	−.106

C.2 Log-Level Uncertainty

Tables 9 and 18 provide the uncertainty estimates for the paired changes in Table 1. TOAD's unsearched and searched arms use the same weights; the original correction family also includes the augmentation-only arm, whose PDMS change is approximately zero. The search arm's log interval includes zero for PDMS, while its geometric changes have mixed directions. The EPDMS intervals for the DiffusionDrive IL-to-RL transition and GTRS-Dense also include zero.

Table 9: **NAVSIM-v1 `navtest`: paired uncertainty.** Holm-adjusted Wilcoxon p-values (first row), scene-bootstrap 95% intervals (second row), and driving-log-bootstrap 95% intervals (third row) for the five v1 quantities of Table 1. Correction uses the original families specified in Appendix A.4, including TOAD's augmentation-only arm.

Pair	PDMS	DDC	Jitter	Smooth.	Adh.
ReCog-2B IL→RL	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.038, .046]$	$[-.015, -.011]$	$[1.15, 1.23]$	$[.180, .185]$	$[.311, .326]$
Log CI	$[.035, .050]$	$[-.021, -.008]$	$[1.07, 1.31]$	$[.176, .189]$	$[.287, .351]$
ReCog-8B IL→RL	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.031, .038]$	$[-.007, -.004]$	$[.88, .95]$	$[.171, .174]$	$[.391, .406]$
Log CI	$[.029, .040]$	$[-.008, -.003]$	$[.84, 1.00]$	$[.168, .177]$	$[.372, .427]$
Qwen-Drive SFT→RL	$<10^{-4}$	0.013	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.023, .029]$	$[-.002, -.000]$	$[.05, .09]$	$[.095, .098]$	$[-.048, -.043]$
Log CI	$[.021, .030]$	$[-.003, .000]$	$[.04, .10]$	$[.093, .101]$	$[-.053, -.038]$
DDv1→v2	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.025, .032]$	$[-.015, -.011]$	$[1.32, 1.41]$	$[.277, .291]$	$[.641, .670]$
Log CI	$[.023, .034]$	$[-.019, -.007]$	$[1.29, 1.45]$	$[.270, .298]$	$[.620, .694]$
NoRD IL→RL	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.101, .114]$	$[.013, .020]$	$[-.24, -.19]$	$[-.034, -.027]$	$[-.007, .013]$
Log CI	$[.096, .120]$	$[.012, .021]$	$[-.26, -.18]$	$[-.036, -.025]$	$[-.013, .019]$
GTRS-Dense expert→reward	$<10^{-4}$	$<10^{-4}$	0.013	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.007, .014]$	$[-.006, -.003]$	$[-.00, .04]$	$[-.002, -.001]$	$[-.065, -.035]$
Log CI	$[.005, .015]$	$[-.006, -.002]$	$[-.02, .06]$	$[-.002, -.000]$	$[-.080, -.018]$
TOAD	0.79	$<10^{-4}$	0.0053	$<10^{-4}$	0.00016
Scene CI	$[.000, .002]$	$[-.005, -.003]$	$[-.16, -.12]$	$[.002, .003]$	$[.005, .010]$
Log CI	$[-.000, .002]$	$[-.006, -.003]$	$[-.17, -.11]$	$[.002, .003]$	$[.004, .011]$

C.3 Qualitative Cases and Search Response

Figure 6 adds six scenes beyond Figure 2, showing oscillatory DiffusionDriveV2 plans that score above both v1 and Human.

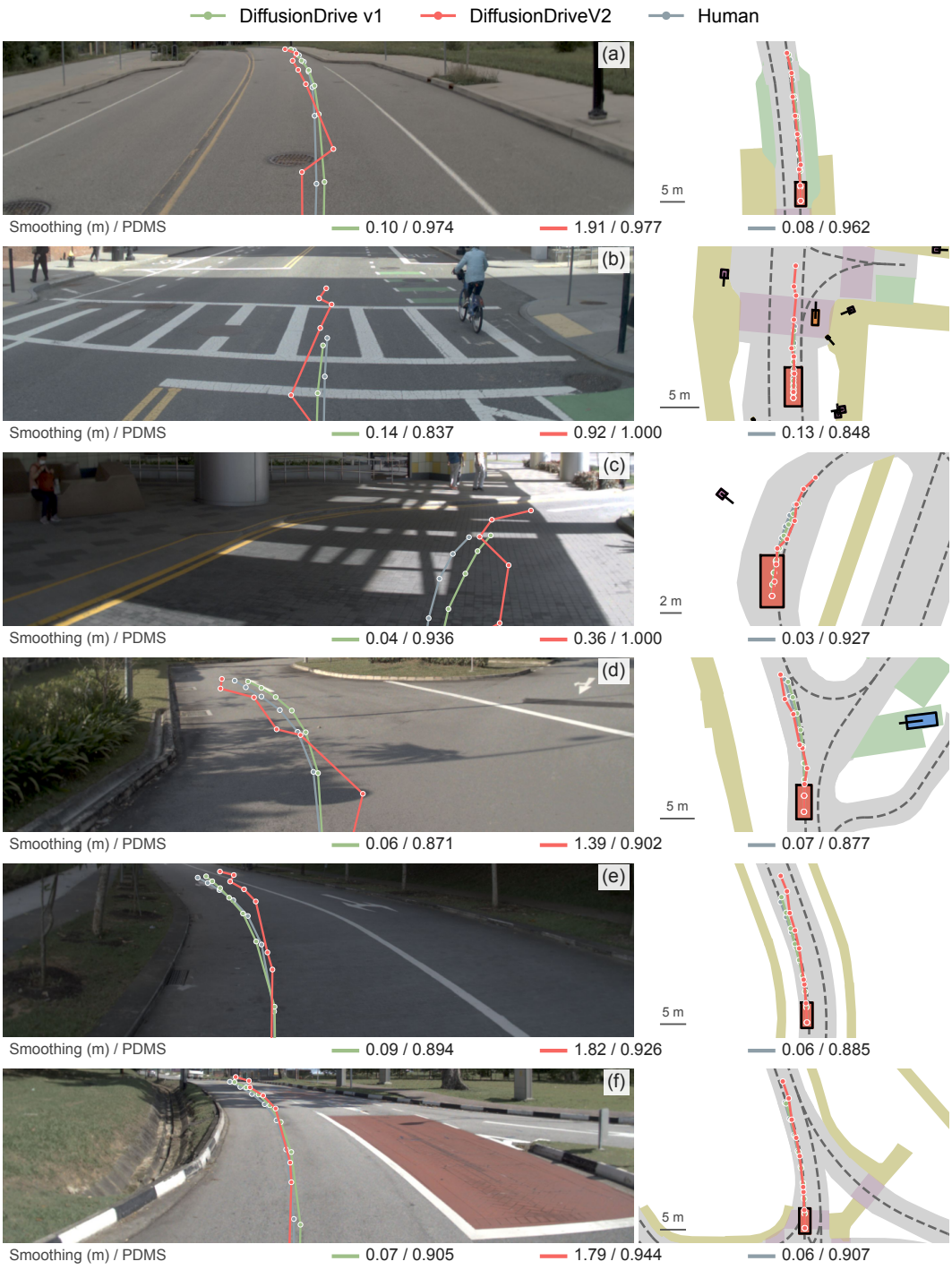

Figure 6: **Additional illustrative trajectory comparisons.** Six selected WorldEngine test scenes (Li et al., 2026a). Each row pairs a cropped front-camera view with a bird's-eye view; curves show planned waypoints. Values give smoothing displacement (m) / PDMS in legend order. In every scene, v2 scores above both v1 and Human while all three pass NC, DAC, TTC, and comfort after tracking. BEV scale bars indicate the per-scene zoom. These cases illustrate the score–geometry mismatch; Table 1 reports aggregate diagnostic changes.

Figure 7 separates TOAD's deployed trajectory from the candidate with the highest predicted score. On `navhard`, that candidate gains 0.008 official PDMS within one CEM round while DDC falls from 0.94 to 0.91; subsequent rounds produce little additional change. For the same candidate sequence, EPDMS falls from 0.809 at round zero to 0.795 after one round. The `navtest` deployed-arm comparison is reported in Tables 1 and 18.

At ten rounds, DDC is 0.931 for the deployed arm and 0.906 for the highest-predicted-score candidate in the sequential trace. TOAD deploys the converged CEM mean subject to its acceptance rule, with comfort and anchoring penalties in the search objective; the comparison evaluates that complete deployment procedure.

Direction under the v2 search objective. We compare the same released TOAD v2 checkpoint without search and with proposal augmentation and CEM (64 samples per round, 8 elites, 5 rounds) (Xu et al., 2026). The predicted-component weights are 10 (NC), 13 (DAC), 6 (DDC), 14 (TTC), 15 (EP), and 2 (comfort). We use the official NAVSIM-v2 `navhard-two-stage` protocol with 450 original scenes and 5,462 synthetic scenes initialized at counterfactual ego states (Cao et al., 2025). Mean deployed DDC changes from 0.9933 to 0.9944 on the original scenes and from 0.9369 to 0.9388 on the synthetic scenes. DDC is included in this target, and its mean does not decline under search in either stage. The v1/v2 comparison evaluates the two released configurations, whose checkpoints and other objective weights also differ.

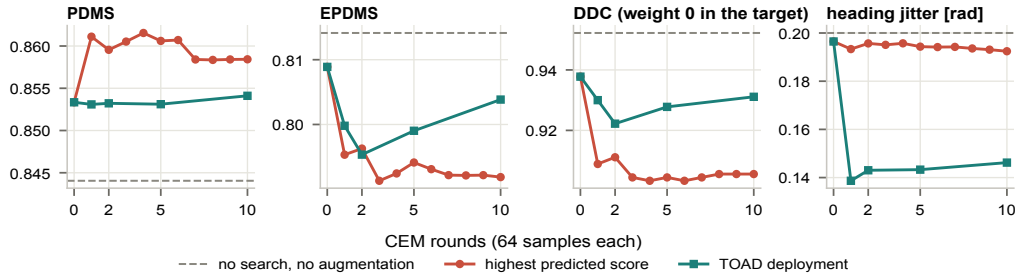

Figure 7: **TOAD search-depth sweep on `navhard`.** Fixed DrivoR weights and 64 samples per CEM round are used throughout. Panels report official PDMS, EPDMS, DDC, and heading jitter (radians). Squares denote TOAD's deployed trajectory; circles denote the candidate with the highest predicted PDMS among all evaluated candidates, starting from the augmented proposal set at round zero. Dashed lines show the planner's argmax without augmentation or search. Candidate ranking uses the learned score; the plotted scores are obtained by evaluating the selected trajectories.

D Closed-Loop Evaluation and Cross-Benchmark Validation

D.1 Full Matched Execution-Interface Comparison

Table 4 evaluates all four conditions on the common scene set per family (1,152 scores), retaining failures. The only experimental switch within each family is the execution interface; checkpoints, scene selection, inference settings, horizon, traffic, and scoring are held fixed. The LQR arm uses NAVSIM's tracker and bicycle motion model, returning the resulting motion state for replanning. Both arms use the same post-hoc scoring configuration. There are 576 identical initial-plan pairs per family, and saved plans, returned frames, and LQR pose consistency were checked. Seeds are fixed; the intervals below describe scene/log variability for released checkpoints, not training or seed variability.

Qwen-Drive and TOAD were run locally; ReCogDrive-2B/8B and NoRD were run on the remote host. Each four-condition comparison is entirely within one host. The two ReCogDrive families use the existing unknown-navigation-command-to-straight adapter in all arms, including the affected scenes. NoRD's full comparison disables prefix caching to restore repeatable initial predictions; the failed attempt was retained and the comparison rerun. These full NR results are distinct from the

earlier playback cohorts below, and are not combined with their open-loop or reactive results to form a new paired contrast.

Table 10: **Paired PDMS gains within each execution interface.** Mean [log-bootstrap 95% CI], the full scene set per row. Interaction estimates and intervals appear in the main comparison (Table 4). Each bootstrap resamples driving logs with all their scenes (2,000 draws, seed 0). Intervals are pointwise.

Planner	ΔS_{PB}	ΔS_{LQR}
ReCogDrive-2B	-0.236 [-0.300, -0.169]	$+0.218$ [+0.165, +0.273]
ReCogDrive-8B	-0.200 [-0.265, -0.132]	$+0.158$ [+0.101, +0.214]
Qwen-Drive	-0.388 [-0.457, -0.321]	$+0.136$ [+0.071, +0.198]
NoRD	$+0.156$ [+0.102, +0.210]	$+0.186$ [+0.128, +0.244]
TOAD	$+0.093$ [+0.043, +0.137]	$+0.073$ [+0.042, +0.104]

Table 11: **Full-cohort component means.** Each entry is reference→optimized, on the same NR scenes per family and interface. NC/DAC are multiplier penalties, EP/TTC/C enter the weighted average, and DDC is diagnostic with zero PDMS weight. PB: playback. TOAD uses Frozen→CEM.

Planner	Interface	NC	DAC	EP	TTC	C	DDC
ReCogDrive-2B	PB	$0.92 \to 0.80$	$0.86 \to 0.67$	$0.59 \to 0.38$	$0.90 \to 0.80$	$1.00 \to 1.00$	$0.97 \to 0.83$
ReCogDrive-2B	LQR	$0.94 \to 0.94$	$0.66 \to 0.91$	$0.57 \to 0.76$	$0.90 \to 0.91$	$1.00 \to 1.00$	$0.93 \to 0.99$
ReCogDrive-8B	PB	$0.93 \to 0.84$	$0.85 \to 0.70$	$0.59 \to 0.40$	$0.91 \to 0.85$	$1.00 \to 1.00$	$0.98 \to 0.85$
ReCogDrive-8B	LQR	$0.95 \to 0.93$	$0.70 \to 0.90$	$0.63 \to 0.76$	$0.91 \to 0.91$	$1.00 \to 1.00$	$0.95 \to 0.99$
Qwen-Drive	PB	$0.96 \to 0.75$	$0.91 \to 0.61$	$0.68 \to 0.40$	$0.96 \to 0.74$	$1.00 \to 0.97$	$0.99 \to 0.86$
Qwen-Drive	LQR	$0.95 \to 0.98$	$0.77 \to 0.91$	$0.69 \to 0.76$	$0.93 \to 0.98$	$1.00 \to 1.00$	$0.99 \to 0.99$
NoRD	PB	$0.92 \to 0.93$	$0.66 \to 0.80$	$0.43 \to 0.61$	$0.91 \to 0.91$	$1.00 \to 1.00$	$0.89 \to 0.97$
NoRD	LQR	$0.92 \to 0.93$	$0.55 \to 0.73$	$0.44 \to 0.64$	$0.89 \to 0.91$	$1.00 \to 1.00$	$0.84 \to 0.95$
TOAD	PB	$0.99 \to 0.99$	$0.86 \to 0.93$	$0.69 \to 0.79$	$0.94 \to 0.96$	$1.00 \to 0.98$	$0.99 \to 0.99$
TOAD	LQR	$0.99 \to 0.99$	$0.92 \to 0.99$	$0.79 \to 0.86$	$0.95 \to 0.97$	$1.00 \to 1.00$	$0.99 \to 1.00$

Qwen speed revisions and alternative explanations. For each of the eight saved plans, vehicle-center positions are obtained from rear-axle poses with the pinned vehicle offset. Longitudinal speed uses the forward difference over 0.1 s projected onto the current heading. Seven same-time revisions are averaged within each scene before aggregation. On the full scene set the means are -0.273 (IL) and $+2.167$ (RL) under playback, versus -0.058 and $+0.584$ under LQR. The paired LQR-minus-playback differences are $+0.215$ $[+0.180, +0.253]$ for IL and -1.584 $[-1.710, -1.453]$ for RL (log-bootstrap 95% intervals). These are descriptive behavioral responses to the interface intervention, not a causal mediation estimate.

Two separately constructed Qwen controls help qualify the interpretation. At fixed original states and non-image inputs, replacing recorded images with renders changes PDMS by -0.079 (IL) and -0.071 (RL); the differential change is $+0.008$ $[-0.037, +0.050]$, on the full paired scene set. This does not establish rendering equivalence at later visited states. Holding the same first predicted plan fixed, playback-minus-LQR changes are $+0.029$ $[-0.001, +0.060]$ for IL and -0.005 $[-0.035, +0.024]$ for RL; the differential change is -0.034 $[-0.075, +0.003]$. These contrasts are not added to or subtracted from the new replanning comparison as a causal decomposition. Together, they motivate examining execution and subsequent feedback jointly rather than attributing the rollout reversal solely to image replacement or single-plan smoothing.

Interpretation and remaining identification limits. The observed Goodhart-related concern is the scope of a score-based improvement claim: optimizing the tracked score need not improve requested-motion diagnostics or preserve gains under another execution interface. Tracking that makes a planner–controller system work well is not itself a benchmark failure. Nor does the observed interaction prove that training deliberately or causally selected a particular discarded component. Qwen-Drive's plan-adherence error improves after RL, and TOAD's increases slightly (Section 3.4); a universal increase in request–execution distance is not supported. NoRD's vocabulary

and TOAD's penalties are possible explanations for their positive gains under both interfaces, not established causal protection mechanisms. Testing those explanations would require matched constraint ablations. The present intervention covers NR traffic and four-second failure-selected scenarios; its conclusions do not establish reactive or real-world transfer.

D.2 Historical Playback Comparisons

Table 12 preserves the earlier playback comparisons in both traffic regimes; Figure 9 shows their NR scene-level distributions. These are historical results, separate from the full matched execution intervention in Table 4. Under this historical playback protocol, ReCogDrive and Qwen-Drive reverse their open-loop ordering, while NoRD retains a gain and TOAD improves despite a small negative open-loop change on the same scenes.

Table 12: **Historical playback gain transfer.** WorldEngine test set (Li et al., 2026a); closed-loop execution uses playback. Open-loop (OL), non-reactive closed-loop (CL-NR), and reactive closed-loop (CL-R) changes. Regimes are matched separately.

	Paired ΔPDMS ↑		
Planner	OL	CL-NR	CL-R
ReCogDrive-2B	**+0.094**	**−0.296**	**−0.248**
ReCogDrive-8B	**+0.057**	**−0.207**	**−0.179**
Qwen-Drive	**+0.050**	**−0.380**	**−0.358**
NoRD	**+0.159**	**+0.087**	**+0.116**
TOAD	**−0.007**	**+0.094**	**+0.060**

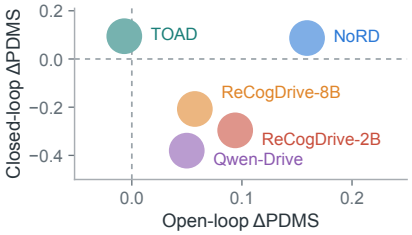

Figure 8: **Historical playback mean gains.** Each circle shows a planner's mean open-loop and non-reactive closed-loop score changes on matched scenes. Colors identify planners; circles have equal size.

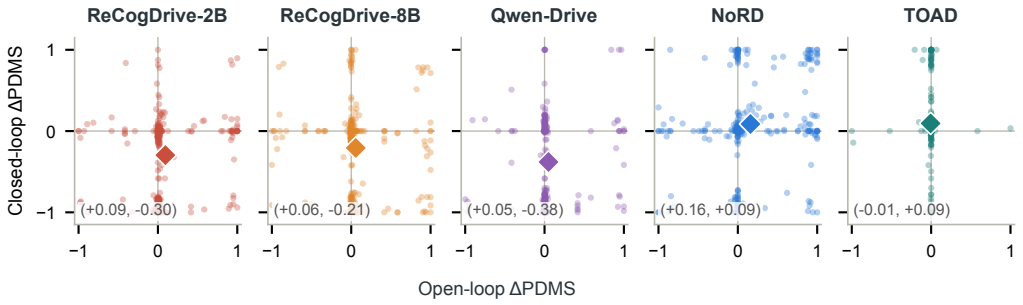

Figure 9: **Historical playback gain transfer on the WorldEngine test set** (Li et al., 2026a). Open-loop versus NR closed-loop PDMS changes (optimized minus baseline): IL→RL for the first four panels, DrivoR→TOAD for the last. Points: scenes; diamonds: paired means. Mean changes appear in Table 12.

In NR evaluation, the counts of scenes with better/worse scores are 62/167 for ReCogDrive-2B, 67/164 for ReCogDrive-8B, 76/153 for Qwen-Drive, 134/83 for NoRD, and 138/51 for TOAD; ties are omitted. Figure 8 uses the same NR-matched scenes and unnormalized PDMS changes. Each equal-size circle is centered on the observed paired mean; marker size does not encode uncertainty.

The NR ReCogDrive changes have paired-test p-values 2×10^{-17} (2B) and 4×10^{-12} (8B); paired contrasts between regimes have $p = 4 \times 10^{-22}$ and 2×10^{-12}. Qwen-Drive's paired scene-bootstrap intervals (seed 0) are $[-0.442, -0.318]$ (NR) and $[-0.417, -0.298]$ (R); its NR-matched open-loop interval is $[0.016, 0.083]$. NoRD's gains have $p = 3 \times 10^{-4}$ (NR) and 10^{-6} (R). TOAD's paired bootstrap intervals are $[0.053, 0.139]$ (NR) and $[0.015, 0.107]$ (R), with $p = 5 \times 10^{-8}$ and 4×10^{-5}.

For ReCogDrive-2B, at-fault collisions rise from 15 to 63 scenes and drivable-area exits from 42 to 99; zero-score rollouts rise from 56 to 150. ReCogDrive-8B's zero-score count rises from 73 to 128. For Qwen-Drive, NR PDMS falls from 0.780 to 0.400: at-fault collisions rise from 11 to 72,

drivable-area exits from 28 to 111, and zero-score rollouts from 39 to 165. With reactive traffic, PDMS falls from 0.766 to 0.408, with collisions increasing from 13 to 57 and drivable-area exits from 28 to 118. NoRD's zero-score count falls from 114 to 94, while TOAD's falls from 46 to 23 (NR) and 45 to 30 (R). TOAD's NR drivable-area exits fall from 41 to 19, while at-fault collisions change from 5 to 4.

Implementation checks. All comparisons use the same adapter. For the verified ReCogDrive pair, the first planning step reproduces the open-loop trajectory within 4 cm for both checkpoints. Re-executing 72 rollouts reproduces all seven score channels bit-identically.

D.3 Successive Predictions and Rollout Diagnostics

Table 13: **ReCogDrive predicted foldbacks.** Historical WorldEngine playback with non-reactive traffic, 286 matched scenes; 0.05 m filter. CL: closed loop. Rates over n predictions (lower is better).

ReCogDrive	Open loop $n = 286$	First CL $n = 286$	All CL $n = 2{,}288$
2B-IL	0.0%	0.0%	17.9%
2B-RL	6.3%	6.3%	36.2%
8B-IL	0.3%	0.3%	7.6%
8B-RL	1.4%	0.7%	35.4%

Qwen-Drive speed revisions. We analyze the complete, matched Qwen-Drive NR rollouts using the stored plans and states. RL increases speed at the first update in 233 scenes and finishes more than 2 m/s above its initial speed in 249; the corresponding IL counts are 101 and 26. For NoRD, the corresponding RL counts are 94 and 33 on its matched set.

Equation 13 in the main text decomposes the observed speed change into the previous plan's intended change and a same-time replanning revision. Speeds are extracted using the official playback conversion, with $h = 0.5$ s; the second term compares predictions at the same absolute future time. Sign counts use a 10^{-8} m/s numerical tolerance. The maximum reconstruction error over the saved records is 3.7×10^{-13} m/s. Across 2,009 adjacent-plan comparisons per checkpoint, the mean revision is -0.262 m/s for IL and $+2.184$ m/s for RL; 576 and 1,775 revisions are positive, respectively. All seven revisions are positive in one IL scene and 196 RL scenes. Meanwhile, 1,897/2,296 RL plans (82.6%) predict a speed decrease from h to $2h$, compared with 1,284 IL plans. Using path length over the same future half-second interval instead of a local velocity difference retains the contrast: mean revisions are -0.205 and $+1.042$ m/s, with all seven positive in 0 and 203 scenes. These are descriptive within-scene sequences; their updates are not treated as independent experimental replicates.

Qwen-Drive case selection. Figure 3 shows WorldEngine NR scene `02b68b9cc51f506a`. We selected the first token in lexical order among 19 scenes with an open-loop gain above 0.01, IL closed-loop score above 0.8, RL closed-loop score zero, initial speed above 2 m/s, and at least seven positive RL speed updates. Across the eight updates, RL speed rises from 4.939 to 17.121 m/s and the vehicle leaves the road at 4 s; its closed-loop PDMS is zero, compared with 0.814 for IL. The full matched-set counts above contextualize this illustrative case.

Continuation to the same junction. For the case in Figure 3, the saved IL continuation extends the rollout to 14 s with 28 half-second decisions in total. Replaying the original predictions reproduces the four-second state sequence exactly; 108 camera files across the two continuation-prefix checks are byte-identical. The continuation uses the original Gaussian assets and later real nuPlan actor records, with fresh planner inference after the original horizon. At 12 s, IL has traveled 37.67 m and is 3.59 m from RL's four-second endpoint (38.32 m traveled); its longitudinal speed is 3.25 m/s versus RL's 17.12 m/s. All four IL footprint corners lie in the official nuPlan drivable area at all 29 saved half-second frames through 14 s; only one RL corner remains inside at 4 s. These map checks describe road keeping at saved states; the original four-second scores remain unchanged, and no standard 14-second PDMS is assigned. The continuation compares behavior near the same

Table 14: **ReCogDrive foldbacks on `navtest` and the WorldEngine test set.** Each cell gives counts at $\epsilon = 0.05\,\text{m}$ / $0.5\,\text{m}$. Column denominators are per configuration. The last column counts scenes with at least one qualifying prediction; the preceding columns count predictions. No outcome-based filtering or new model inference is used.

		`navtest`	WorldEngine test set (NR)			
Planner	Variant	Full OL	Matched OL	First CL	All CL	Any CL
	Denominator	12,146	286	286	2,288	286
ReCogDrive-2B	IL	14 / 0	0 / 0	0 / 0	410 / 13	146 / 7
	RL	522 / 31	18 / 2	18 / 1	829 / 528	226 / 197
ReCogDrive-8B	IL	13 / 0	1 / 0	1 / 0	174 / 1	68 / 1
	RL	64 / 0	4 / 0	2 / 0	810 / 646	233 / 214

junction with different arrival times and speeds. Figure 3 samples each row chronologically, using independent time grids to retain the continuation in the IL row.

ReCogDrive prediction foldbacks. We read native eight-point open-loop trajectories and the corresponding half-second knots of interpolated closed-loop plans. Each point is relative to the current ego rear axle. Define a prefix foldback by

$$I_\epsilon(u) = \mathbf{1}[x(h) > \epsilon \ \wedge \ x(h) - x(2h) > \epsilon]\,. \tag{27}$$

We use $\epsilon = 0.05\,\text{m}$ to exclude small displacements and also report $0.5\,\text{m}$; these are descriptive amplitude filters, not safety boundaries. Projecting the second displacement onto the predicted heading at h retains every counted event at its respective threshold. All four checkpoints have complete open-loop outputs on `navtest`. For each NR matched set, we count the eight predictions used for state updates, excluding a terminal prediction that is saved but not executed. Table 14 reports all configurations and both filters, including IL increases. The matched-token open-loop and initial closed-loop records come from their respective runs; matching scene identity does not make their observations or predictions identical.

D.4 nuPlan Score Reproduction

Plan-R1's reinforcement scores agree with the published CLS values within 0.1 points in both regimes (Table 15). Recomputing scenario scores from stored sub-metrics agrees with the official aggregate to 10^{-15}.

Table 15: **nuPlan `val14`: score reproduction.** CLS in score points ($\times 100$); a dash denotes no published value.

	Non-reactive		Reactive	
Checkpoint	Published ↑	Ours ↑	Published ↑	Ours ↑
PDM-Closed	92.84	92.84	92.12	92.13
Plan-R1 imitation	–	85.48	–	81.01
Plan-R1 reinforcement	88.98	88.89	87.69	87.71

D.5 nuPlan Paired Scores and Motion

Tables 16 and 17 retain all official score components and the motion diagnostics used in the main analysis. Both traffic settings use the same `val14` scenarios; each retains the original correction family defined in Appendix A.4. The score, progress, and plan-consistency improvements are reported alongside the increased absolute jerk peaks; unsupported changes remain identified rather than interpreted as effects.

Table 16: **nuPlan `val14`: non-reactive traffic (NR).** Plan-R1 imitation and reinforcement means, paired changes and uncertainty. † does not pass all three statistical checks of Appendix A.4.

Quantity	IL	RL	Δ	Scene CI	Log CI	p_{Holm}
CLS ↑	0.855	0.889	+0.034	[0.018, 0.050]	[0.019, 0.050]	$< 10^{-4}$
Drivable-area compliance ↑	0.942	0.984	+0.042	[0.030, 0.054]	[0.030, 0.056]	$< 10^{-4}$
No-at-fault-collision gate ↑	0.956	0.953	−0.004†	[−0.016, 0.008]	[−0.015, 0.008]	1.000
Driving-direction gate ↑	0.989	0.980	−0.009†	[−0.016, −0.002]	[−0.017, −0.002]	0.134
Making-progress gate ↑	0.986	0.999	+0.013	[0.007, 0.021]	[0.007, 0.020]	0.002
Progress term ↑	0.908	0.923	+0.015	[0.007, 0.023]	[0.006, 0.023]	0.022
TTC term ↑	0.901	0.902	+0.001†	[−0.015, 0.018]	[−0.017, 0.018]	1.000
Speed-limit compliance ↑	0.980	0.996	+0.017	[0.013, 0.021]	[0.013, 0.020]	$< 10^{-4}$
Comfort term ↑	0.999	0.997	−0.002†	[−0.005, 0.002]	[−0.006, 0.002]	1.000
Absolute lon. jerk peak [m/s^3] ↓	0.765	0.901	+0.136	[0.110, 0.162]	[0.107, 0.168]	$< 10^{-4}$
Absolute jerk peak [m/s^3] ↓	0.773	0.841	+0.068	[0.043, 0.093]	[0.043, 0.092]	$< 10^{-4}$
Planned heading jitter ↓	1.780	1.546	−0.235	[−0.272, −0.199]	[−0.280, −0.193]	$< 10^{-4}$
Smoothing displacement [m] ↓	0.068	0.068	−0.000†	[−0.001, 0.001]	[−0.001, 0.001]	0.344
Plan churn [m] ↓	0.418	0.382	−0.037	[−0.044, −0.030]	[−0.044, −0.030]	$< 10^{-4}$
First-plan open-loop score ↑	0.753	0.773	+0.020	[0.008, 0.032]	[0.007, 0.032]	$< 10^{-4}$

Table 17: **nuPlan `val14`: reactive traffic (R).** Plan-R1 imitation and reinforcement means, paired changes and uncertainty. † does not pass all three statistical checks of Appendix A.4.

Quantity	IL	RL	Δ	Scene CI	Log CI	p_{Holm}
CLS ↑	0.810	0.877	+0.067	[0.051, 0.083]	[0.052, 0.083]	$< 10^{-4}$
Drivable-area compliance ↑	0.948	0.981	+0.033	[0.022, 0.045]	[0.021, 0.047]	$< 10^{-4}$
No-at-fault-collision gate ↑	0.934	0.944	+0.010†	[−0.003, 0.022]	[−0.002, 0.021]	0.598
Driving-direction gate ↑	0.988	0.983	−0.005†	[−0.011, 0.001]	[−0.012, 0.001]	0.598
Making-progress gate ↑	0.960	0.996	+0.037	[0.026, 0.048]	[0.026, 0.047]	$< 10^{-4}$
Progress term ↑	0.851	0.895	+0.044	[0.033, 0.055]	[0.034, 0.055]	$< 10^{-4}$
TTC term ↑	0.888	0.902	+0.013†	[−0.003, 0.029]	[−0.004, 0.031]	0.598
Speed-limit compliance ↑	0.982	0.996	+0.014	[0.011, 0.019]	[0.011, 0.018]	$< 10^{-4}$
Comfort term ↑	0.996	0.987	−0.009	[−0.015, −0.004]	[−0.015, −0.004]	0.039
Absolute lon. jerk peak [m/s^3] ↓	0.845	1.051	+0.206	[0.176, 0.235]	[0.172, 0.242]	$< 10^{-4}$
Absolute jerk peak [m/s^3] ↓	0.812	0.957	+0.145	[0.119, 0.171]	[0.118, 0.173]	$< 10^{-4}$
Planned heading jitter ↓	1.437	0.899	−0.538	[−0.610, −0.471]	[−0.612, −0.463]	$< 10^{-4}$
Smoothing displacement [m] ↓	0.068	0.071	+0.003	[0.002, 0.004]	[0.002, 0.004]	$< 10^{-4}$
Plan churn [m] ↓	0.467	0.448	−0.019	[−0.029, −0.008]	[−0.030, −0.007]	$< 10^{-4}$
First-plan open-loop score ↑	0.745	0.770	+0.025	[0.013, 0.036]	[0.011, 0.037]	$< 10^{-4}$

E EPDMS Rescoring and Two-Stage Evaluation

E.1 Fixed-Output EPDMS Rescoring

Table 18 collects the scene- and log-level intervals and adjusted tests for the fixed-output `navtest` changes in Table 1. These single-stage readouts use v2 direction, lane keeping, and history comfort; v2 DDC excludes intersection positions and enters the score as a gate. The cross-frame and stage aggregation of the full two-stage experiment is evaluated separately below.

Table 8 in Appendix C.1 reports EPDMS rescoring on `navhard` and the WorldEngine test set alongside the original PDMS and diagnostic changes. NoRD improves the revised score and displayed components on both sets. ReCogDrive retains score gains, while some component changes have weaker support on the smaller sets. For Qwen-Drive, the `navtest` lane-keeping and history-comfort declines are not significant on either smaller set. These are evaluations of fixed outputs, separate from optimizing the revised objective.

Table 18: **EPDMS rescoring on `navtest`: paired statistical support.** For each pair, rows give Holm-adjusted Wilcoxon p-values, scene-bootstrap 95% intervals, and driving-log-bootstrap 95% intervals for the mean changes reported in Table 1. PDMS uses the v1 scorer; EPDMS, driving-direction compliance (DDC), lane keeping (LK), and history comfort (HC) use the v2 scorer on fixed outputs. Holm correction retains the original test families in Appendix A.4, including TOAD's augmentation-only arm, without recomputing tests for the displayed subset.

	NAVSIM-v1	NAVSIM-v2 scorer			
Pair	ΔPDMS	ΔEPDMS	ΔDDC	ΔLK	ΔHC
ReCog-2B IL→RL	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.038, .046]$	$[.015, .024]$	$[-.012, -.009]$	$[-.063, -.052]$	$[-.007, -.004]$
Log CI	$[.035, .050]$	$[.009, .029]$	$[-.017, -.006]$	$[-.073, -.043]$	$[-.008, -.004]$
ReCog-8B IL→RL	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.031, .038]$	$[.013, .021]$	$[-.006, -.004]$	$[-.029, -.020]$	$[-.005, -.003]$
Log CI	$[.029, .041]$	$[.010, .024]$	$[-.007, -.003]$	$[-.037, -.014]$	$[-.006, -.002]$
Qwen-Drive SFT→RL	$<10^{-4}$	$<10^{-4}$	0.00083	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.023, .029]$	$[.016, .021]$	$[-.002, -.000]$	$[-.016, -.008]$	$[-.006, -.003]$
Log CI	$[.021, .030]$	$[.014, .023]$	$[-.002, -.000]$	$[-.019, -.004]$	$[-.007, -.002]$
DDv1→v2	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.025, .032]$	$[-.001, .007]$	$[-.011, -.008]$	$[-.024, -.016]$	$[-.098, -.087]$
Log CI	$[.023, .035]$	$[-.005, .011]$	$[-.015, -.005]$	$[-.028, -.012]$	$[-.107, -.079]$
NoRD IL→RL	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.101, .114]$	$[.091, .104]$	$[.012, .017]$	$[.008, .019]$	$[.017, .024]$
Log CI	$[.096, .120]$	$[.086, .109]$	$[.010, .019]$	$[.005, .023]$	$[.017, .025]$
GTRS-Dense expert→reward	$<10^{-4}$	$<10^{-4}$	0.034	$<10^{-4}$	$<10^{-4}$
Scene CI	$[.007, .014]$	$[-.001, .007]$	$[-.002, -.000]$	$[-.013, -.005]$	$[.006, .010]$
Log CI	$[.005, .016]$	$[-.002, .008]$	$[-.002, .000]$	$[-.014, -.004]$	$[.004, .013]$
DrivoR→TOAD	0.79	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.42
Scene CI	$[.000, .002]$	$[-.004, -.002]$	$[-.004, -.002]$	$[-.010, -.006]$	$[-.002, .000]$
Log CI	$[-.000, .002]$	$[-.005, -.001]$	$[-.005, -.001]$	$[-.012, -.005]$	$[-.002, .000]$

E.2 Reinforcement Pairs under the Two-Stage Protocol

We run the official NAVSIM-v2 `navhard-two-stage` protocol on 450 original scenes and 5,462 synthetic scenes for the ReCogDrive, Qwen-Drive, and NoRD pairs. The protocol is a re-inference, not a rescoring: the 5,462 synthetic scenes start from counterfactual ego states and are rendered from those states, so each planner produces new trajectories from the rendered eight-camera observations (the synthetic assets contain no LiDAR, which excludes DiffusionDrive and DiffusionDriveV2, whose feature extractors consume the point cloud). The planners run outside the v2 evaluation code through the same adapters as the closed-loop experiments: Qwen-Drive receives the three camera views at the four most recent frames and a history interpolated from the four ego states, ReCogDrive receives the front frame and the four ego statuses, and NoRD runs its released agent against its served model; on fourteen synthetic scenes whose navigation command is undefined, ReCogDrive is given the straight-ahead command that NoRD's own agent substitutes in the same case. Predictions are then scored and aggregated by the unmodified official code, with the mapping between original and synthetic scenes taken over the full split. Table 3 reports scene-mean components for each stage and the official combined EPDMS. Paired bootstrap 95% intervals resample the 450 original scenes for stage one and the 450 official mapping groups for stage two and the combined score. The stage-two score changes below use group-weighted means, which can differ from unweighted scene means.

The three free-form pairs gain 0.105, 0.079, and 0.063 PDMS on the same 450 scenes in open loop and 0.077, 0.060, and 0.062 EPDMS under fixed-output rescoring (Table 8). Under the two-stage protocol their stage-one changes are $+0.000$, -0.015, and -0.003: the original scenes are re-planned from rendered rather than recorded frames, and the open-loop advantage does not survive the re-inference. On the synthetic stage the changes are -0.012, -0.016, and -0.010, with intervals containing zero. The combined changes are -0.008 $[-0.038, +0.019]$, -0.008 $[-0.035, +0.018]$, and $+0.003$ $[-0.021, +0.027]$, respectively; these intervals include zero and are marked † in Table 3. NoRD gains $+0.143$ and $+0.049$ in stages one and two, and $+0.079$ $[+0.048, +0.112]$ in combined EPDMS.

Table 19 decomposes the synthetic stage. For both ReCogDrive pairs, DAC and EP increase while DDC decreases; Qwen-Drive improves all displayed components except EC. The term that moves most is two-frame extended comfort. Within each stage, it compares time-aligned simulated motion from consecutive paired frames. On the synthetic stage it falls by 0.378, 0.527, and 0.296, from 0.65–0.69 to 0.15–0.35, while NoRD's rises by 0.038. The reinforcement checkpoints therefore pass the cross-frame motion-consistency checks less often, complementing the closed-loop prediction-revision measurements in Appendix D.3.

For original-scene scores $S_{1,g}$ and weighted synthetic-child scores $\bar{S}_{2,g}$, the official combined score is $\mathbb{E}_g[S_{1,g}\bar{S}_{2,g}]$ rather than $\mathbb{E}_g[S_{1,g}]\mathbb{E}_g[\bar{S}_{2,g}]$, so changes in the two stage means need not share the direction of the combined change.

Table 19: **NAVSIM-v2 `navhard-two-stage`: component analysis.** Synthetic-stage paired changes of the EPDMS components. EC is the two-frame extended comfort.

	Stage 2 component changes ↑							
Pair	ΔNC	ΔDAC	ΔDDC	ΔEP	ΔTTC	ΔLK	ΔHC	ΔEC
ReCog-2B IL→RL	$-.001$	$+.005$	$-.011$	$+.061$	$-.001$	$+.006$	$-.000$	$-.378$
ReCog-8B IL→RL	$-.000$	$+.008$	$-.002$	$+.056$	$-.003$	$+.003$	$-.002$	$-.527$
Qwen-Drive SFT→RL	$+.005$	$+.029$	$+.008$	$+.006$	$+.005$	$+.005$	$+.008$	$-.296$
NoRD SFT→RL (vocab.)	$+.024$	$+.120$	$+.051$	$+.040$	$+.022$	$+.024$	$+.017$	$+.038$