DRIVING COMPETENCE IS NOT A SCALAR: A PARETO AUDIT AND PROFILE FOR END-TO-END DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

A single number ranks end-to-end driving planners, and NAVSIM's PDMS (Dauner et al., 2024; Cao et al., 2025), built to fix its predecessor, is that number. Each repair so far has adjusted weights inside the number and kept its type. We argue the type is wrong. Driving competence rests on attributes that genuinely trade off, so the object of scoring is a partial order rather than a quantity. On `navtest`, under a four-axis operationalization whose partition we also sweep, 62–71% of admissible trajectory pairs are Pareto-incomparable, and the per-scene competence poset has certified Dushnik–Miller order dimension ≥ 3 in $\sim$91% of scenes and dimension 1 in none. A scalar's own order is total, so on every such pair it must either collapse a real distinction or order it by an undeclared preference, and that contested set is exactly the incomparable set at any positive weighting. Known complaints follow. PDMS separates a human from a strong sensor-based planner by under 0.01 in 46.3% of scenes, its per-scene winner contradicts its aggregate ranking in up to 28.8% of scenes, and on a deployed pair that closed-loop testing separates by six points, the sign of its verdict turns on scorer implementation and weighting. The Driving Competence Profile (DCP) reports competence at its type. It returns an admissibility gate, four graded axes calibrated independently of the evaluated planners, and a leaderboard scalar carried with the profile and a map of which comparisons it arbitrates. Over 13 zoo planners plus current state-of-the-art methods, it ranks the human above a blind no-perception model in 83.4% of scenes against the official PDMS's 71.6%, and it flags evidence consistent with an external Goodhart step, a method whose published PDMS rose $88.1 \rightarrow 91.2$ while its measured competence fell.

1 INTRODUCTION

The number that ranks end-to-end driving today exists to fix the last one. Open-loop L_2 on nuScenes (Caesar et al., 2020) fell when perception-free models matched full stacks on it (Zhai et al., 2023; Li et al., 2024), and the field moved to simulation-aligned scoring. NAVSIM's Predictive Driver Model Score (PDMS) (Dauner et al., 2024) and its reactive two-stage extension EPDMS (Cao et al., 2025) simulate the trajectory and aggregate collision, compliance, progress, time-to-collision and comfort (lane keeping since v2) into one score.

That repair worked in aggregate. Leaderboards (Hu et al., 2023; Jiang et al., 2023; Liao et al., 2025) cluster within a few points, and a cross-benchmark study finds PDMS tracking closed-loop driving at the method level (Wang et al., 2026). The correlation is the metric's best defence and also its puzzle, a number right about methods in aggregate and wrong on the individual pairs a leaderboard is read for. The sequence of repairs repeats one shape. L_2 on nuScenes fell to the ego-status exploit and gave way to PDMS; PDMS saturated and rewarded aggression and gave way to EPDMS; under EPDMS the ego-status shortcut returns (§3). Each step retired a metric by an exploit that metric could not see, and each successor is a single number again (App. A), re-weighted *inside* but of the same type.

We argue the issue is one of type. Driving competence rests on attributes that genuinely trade off, safer-but-slower against faster-but-closer, so as a measurement problem (Stevens, 1946; Krantz et al., 1971) its construct is a *partial order*. A score is a map into $(\mathbb{R}, \leq)$, which has no incomparable

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.

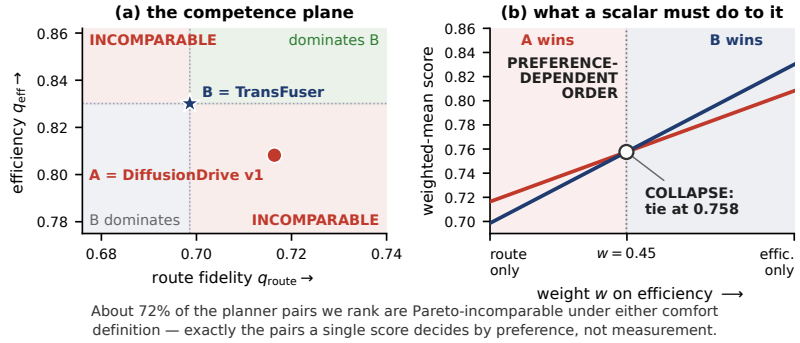


Figure 1: **Competence is a partial order, and no scalar represents one.** (a) DiffusionDrive v1 (A) and TransFuser (B) are incomparable. (b) A weighted-mean sweep across that pair changes which planner is preferred, then collapses it to a tie (0.758).

elements, and no choice of sub-metrics or weights repairs that mismatch (Fig. 1a). The documented pathologies of PDMS are then shadows of one fact rather than independent bugs. Optimizing a proxy invites Goodhart's law (Strathern, 1997), and the shortcut PDMS set out to defeat still scores ≈ human on it (§3).

Contributions.

1. **A measured premise** (§2): competence is empirically a partial order, and on one deployed pair that closed-loop testing separates by six points, PDMS's verdict flips with scorer implementation and weighting.
2. **The pathologies as corollaries** (§3): what practitioners report about PDMS is what that premise forces; we quantify four (P7 new) plus a fifth failure of purpose, and show EPDMS inherits them.
3. **The DCP diagnostic protocol** (§4): an admissibility gate and a graded Pareto vector of four axes calibrated independently of the evaluated planners, reported with a leaderboard scalar *and* the map of which comparisons it arbitrates.
4. **A validation battery** over 13 zoo planners plus current SOTA (§5): evidence consistent with a Goodhart step by other researchers; the most-cited closed-loop inversion, corrected on its pair (the three-way order is a preference call, §5.2); and **a released scoring interface**.

2 Competence is a partial order

Three metrics and three patches later, doing better than suspicion requires saying what we measure, so we define the object first and then measure how wide it is.

Definition 1 (Competence preorder)**.** Fix a scene s with assigned route. A trajectory τ over the 4 s horizon has a competence attribute vector $q(\tau) = (q_{\text{saf}}, q_{\text{cmf}}, q_{\text{route}}, q_{\text{eff}}) \in [0,1]^4$ (§4), defined only on the *admissible* set $\mathcal{A}_s$ (legally on-route). Define $\tau \preceq \tau'$ iff $q_k(\tau) \le q_k(\tau')$ for all k. Then $\preceq$ is a preorder on $\mathcal{A}_s$, and a partial order on distinct attribute vectors; τ, τ' are *incomparable* if neither $\tau \preceq \tau'$ nor $\tau' \preceq \tau$. Below we use only the completeness $\preceq$ lacks.

A scalar metric is a map $m : \mathcal{A}_s \to \mathbb{R}$ factoring through the attribute vector, $m = f \circ q$, and it induces the total preorder $\tau \preceq_m \tau' \iff m(\tau) \le m(\tau')$. Call m *Pareto-monotone* if $\tau \prec \tau' \Rightarrow m(\tau) < m(\tau')$, as every driving-leaderboard metric is (the PDMS product, weighted sums, power means). Whether such an m can agree with $\preceq$ on every pair is decided by the structure of $\preceq$ (Stevens, 1946; Krantz et al., 1971). We say m *mis-ranks* a pair when it orders or ties a pair $\preceq$ leaves incomparable. The term is descriptive rather than pejorative. A scalar *with* a declared preference completes $\preceq$ legitimately, as DCP's own does (Alg. 1, line 4); the counts below quantify what that completion decides, and PDMS declares none.

All of this would be idle if trajectories rarely traded off, so before proving anything we measure the width of $\preceq$ on 484 `navtest` scenes holding 10,386 admissible candidates. Comfort admits two operationalizations, the shipped seam-neutral one (B) and the ego-seeded one it replaces (A). We carry both, A first, and adopt B on a dose–response test external to the axes (Table 4, App. E).

One scene makes the abstraction concrete. On `c090cc34`, TransFuser and DiffusionDrive v1 score $(1.000, 0.841, 0.523, 0.768)$ and $(1.000, 0.811, 0.693, 0.290)$ on safety, comfort, route fidelity and efficiency. One holds the route and barely moves, the other makes progress and drifts off it. Which drove better? Nothing in the data compels an answer, yet our own declared scalar answers anyway, 0.762 against 0.635, on a weighting nobody voted on.

Measurement 1 (Width of the competence relation)**.** On the pool, between 62% (A) and 71% (B) of admissible trajectory pairs are Pareto-incomparable, so neither trajectory is at least as good on all four axes.

Rescaling an axis cannot change this, since dominance depends only on the per-axis order. Re-operationalizing all four leaves it between 56% and 87% over 256 combinations (App. C), and four *independent* axes would leave 87.5% incomparable, so the axes are coupled rather than inflated.

How many scalars it would take is a question correlations answer weakly, so we ask it order-theoretically (App. C). The Dushnik–Miller *order dimension* (Dushnik & Miller, 1941) of a poset is the least number of total orders whose intersection recovers it. Here $\dim \geq 2$ says no single scalar represents the relation, $\dim \geq 3$ says two do not either, and $\dim \leq 2$ is decidable exactly per scene (Golumbic, 1980).

Measurement 2 (Order dimension of the per-scene poset)**.** $\dim \geq 3$ in about 91% of `navtest` scenes and $\dim = 1$ in none. The exact value beyond 3 we do not claim.

The floor barely moves (91% ego-seeded A, 90.5% seam-neutral B, 79.8–92.1% across ten variants), which is why the two-scalar claim rests here rather than on correlations.

The planners behave like the candidate pool. About 72% of per-scene pairs among the eleven non-human planners we rank are Pareto-incomparable under either comfort definition, and TransFuser and DiffusionDrive v1 are incomparable on 71.8% of the scenes where both are admissible. The fraction stays above half when pairs differing by less than 10^{-2} per axis count as ties, and the weight-free Dilworth bound alone puts a third of all pairs beyond any scalar's reach (Table 8).

The premise is therefore non-vacuous. A single incomparable pair already defeats faithfulness, because $(\mathbb{R}, \leq)$ is total while m must assert $<$, $=$ or $>$ on a pair $\preceq$ leaves unrelated (Lemma 1). The operative question is quantitative. Which comparisons must a scalar get wrong, and how many?

Proposition 1 (The misrepresented set is the incomparable set)**.** *Let I_s be the incomparable fraction of the $\binom{N}{2}$ candidate pairs in scene s, and P_{inc} the incomparable pairs.*

- ***(a)*** Any *scalar m misrepresents $\geq I_s$ of the candidate pairs; for a Pareto-monotone m the misrepresented set is* exactly *P_{inc}, split between falsely-ordered and collapsed pairs, $\mathrm{inv}(m) + \mathrm{col}(m) = |P_{\mathrm{inc}}|$.*
- ***(b)*** *The set on which two legitimate scalars (Pareto-monotone, containing the positive-linear family) can disagree is* exactly *P_{inc}; no comparable pair is contested, every incomparable pair is.*
- ***(c)*** *A weight-free, axis-free antichain-width bound holds, $I_s \geq \binom{w_s}{2} / \binom{N}{2}$, with w_s the poset width (Dilworth, 1950).*

Proofs are in App. B; there are only two ways for a scalar to go wrong.

Corollary 1 (Invert-or-collapse)**.** *On every incomparable pair a scalar m either* strictly orders *it, committing to one of many equally valid weightings, another of which reverses it (b), or* ties *it, collapsing a real distinction. There is no preference-free escape.*

Both horns are visible on real data. The weight sweep of Fig. 1b inverts the DiffusionDrive v1/TransFuser ranking and then collapses it to a tie, and PDMS shows the collapse horn at scale, tying the human, TransFuser and a no-perception baseline in 13.0% of scenes (§3). Of the pairs a scalar ties to within 0.02, 99.9% are distinct vectors. Re-weighting cannot recover what that collapse destroys, because Proposition 1(a) caps per-pair faithfulness at the comparable fraction $1 - I_s$; a dense power-mean family saturates the ceiling exactly and a monotone learned regressor falls just short of it (App. C).

Remark 1 (The profile is a multi-utility representation)**.** A preorder admits a single-utility representation only if it is complete (Debreu, 1954); in general a *set U* of utilities represents it, with

$\tau \preceq \tau' \iff u(\tau) \leq u(\tau')\ \forall u \in U$ (Ok, 2002; Evren & Ok, 2011). Vectorizing is what an incomplete preorder admits, and App. L checks that the axis choice does not install the incompleteness. The choice left is *which* attributes enter U (§4).

2.1 THE ARGUMENT DOES NOT NEED OUR CANDIDATE POOL

The argument's sharpest form uses two planners other people deployed. Closed-loop evaluation separates DriveSuprim and SparseDriveV2 by 6.1 points (83.02 against 89.15, §5.2), and the published PDMS calls the pair the *other* way (93.5 against 92.0). On our reproductions the verdict is not stable enough to call either way. The official devkit puts SparseDriveV2 ahead by $+0.0092$, inside the near-tie band that P7 quantifies (§3), and the sign flips back under our coarser rescoring (-0.018, App. D).

Nor do the weights anchor it. PDMS is a compliance gate times a weighted mean, so the gap reachable over all positive weightings is the convex hull of the per-sub-metric gated-mean differences, and under the official scorer that hull spans zero ($[-0.0415, +0.0920]$, App. D). Its positive mass sits on two-frame comfort and on ego progress, the axis P5 shows to be aggression-biased (§3). A pair the closed-loop referent separates by six points is therefore one the scalar decides by scorer implementation and weight choice, which is how a metric can track closed-loop driving in aggregate and still drop individual pairs (Wang et al., 2026).

3 THE PATHOLOGIES ARE COROLLARIES

If the failure is structural, the reported oddities of PDMS should look like one fact seen from several sides rather than a list of unrelated bugs. Table 5 (App. F) quantifies four, numbered as in the full enumeration there. Three of them share a cause, which is that scale spent certifying legality has almost none left to grade driving. The compliance gate carries 82–84% of $\mathrm{Var}(\text{PDMS})$ on the three full-`navtest` runs we hold (P1; median 77% across the 42 subset runs, 94–98% over the 14 learned planners under the official scorer). Inside the compliant band, which holds 95.7% of scenes, grading nearly vanishes (P2; official scorer 73–81%, ours 9–11%). P5, aggression bias, says which way the little grading that remains points, since EP rises with distance travelled and so ranks the prudent human below a more aggressive compliant anchor in 99.8% of scenes.

P7 is new here, and it is the per-scene face of P2. Under the official scorer PDMS separates the human from TransFuser by less than 0.01 in 46.3% of scenes, and it ties human, TransFuser and a no-perception baseline exactly in 13.0%. The aggregate order human>TF>Ego is genuine but *emergent*. On the two close pairs the metric's own per-scene winner reproduces that order on under half the scenes (43.7–48.8%) and contradicts it on up to 28.8%, and only the widest pair (human–Ego) reaches 63.6% (median pairwise separation 0.0124). Where planners are close, which is where people consult a leaderboard, PDMS cannot support the per-scene attribution it is increasingly used for (Figs. 5 and 6, App. F).

Those four concern how PDMS grades. The fifth is a failure of its purpose. PDMS exists because ego-status-only models gamed open-loop L_2 (Zhai et al., 2023; Li et al., 2024). Yet under the benchmark's own reported protocol, whose human-penalty filter is itself human-referenced, PDMS ranks the human above our blind EgoStatusMLP reproductions in only 71.6% of scenes on the 1098-token split, against DCP's 83.4% with disjoint intervals (App. I). DCP separates them because its route axis grades precisely what a blind model cannot do, follow the road.

EPDMS removes none of these pathologies. Where we can re-measure under reactive agents we do, and 25.7% of scenes are still exact ties on our in-house scorer. Where we cannot, the structure decides it, because the extra two-frame term is the sign-carrying term of the fragile pair in §2.1 (App. D). What has to change is what is gated, graded and reported.

4 THE DRIVING COMPETENCE PROFILE

Measurement theory prescribes the remedy. Measure the construct at its true type, a vector ordered by Pareto dominance, and scalarize only under an explicit preference. Because $\preceq$ is defined only on $\mathcal{A}_s$, admissibility is a precondition for being scored rather than a fifth attribute, and inside $\mathcal{A}_s$ the

Table 1: **Where the three schemes differ:** what is gated, graded and reported. PDMS_{dk} = official one-stage v2 score (App. A); EPDMS = two-stage, reactive (Cao et al., 2025). *Binary weight* counts weighted-mean mass on per-scene pass/fail terms; EPDMS ties are our reactive re-measure.

	PDMS_{dk}	EPDMS	**DCP**
Compliance enters as	mult. gate	mult. gate	categorical veto
Graded terms	5 sub-metrics	5 sub-metrics	4 axes
Aggregation	weighted mean	weighted mean, 2-stage	power mean S_p
Reported	scalar	scalar	scalar, profile, map
Binary weight	$11/16$ wt. units	$11/16$ wt. units	safety axis (declared)
Human vs. blind: exact ties	13.0% scenes	25.7% scenes	—
Human ranked above blind	71.6%	—	$\mathbf{83}\%$
Deployed pair sign weight-stable?	no	no	n/a (adds axes)
Incomparable pairs	ordered or tied	ordered or tied	reported (62–71%)

profile is the multi-utility representation an incomplete preorder admits (Remark 1). DCP's two tiers are that split (Table 1).

Tier 1: admissibility gate (categorical). Being on drivable area and inside the route lane is genuinely pass/fail, and those two checks are what the released gate implements. A failing trajectory gets no competence score that could be averaged upward. Keeping legality categorical is the direct answer to P1, because the gate keeps its veto but can no longer carry the $\sim 80\%$ of variance that should be grading driving.

Tier 2: competence profile (graded). Remark 1 left one thing to a modelling choice, which attributes enter U, and here we make it. The four axes are the constituents PDMS already presumes and then averages away, each mapped into $[0, 1]$ through its own calibration (Fig. 2) and each answering a specific failure. *Route fidelity* pairs cross-track error with a heading-vs-tangent term, rewarding steering as the road requires where $\sum|\Delta\text{heading}|$ rewards never steering. *Efficiency* is two-sided about the human's pace, which removes the one-sidedness of EP. *Safety* grades time-to-infraction, following RSS and nuPlan (Shalev-Shwartz et al., 2017; Caesar et al., 2021), so a near-miss no longer scores like a clear margin. *Comfort* softens jerk, lateral acceleration and yaw rate against the nuPlan/ISO 2631 envelope (ISO, 1997).

Let $\sigma(x;\theta) = 1/(1+(x/\theta)^2)$ be the soft-saturation kernel ($\sigma(\theta) = 0.5$, monotone decreasing). For trajectory τ with horizon $H{=}4\,\text{s}$:

$$\textbf{Gate: } g = \mathbf{1}[\text{ego corners on drivable area}] \cdot \mathbf{1}[\max_t \text{xt}(\tau_t) \le 1.75\,\text{m}];$$

$$\textbf{Safety: } q_{\text{saf}} = \min\big(\text{clip}(t_{\text{col}}/H),\, \text{clip}(t_{\text{ttc}}/H)\big),\;\; t_\bullet{=}\infty \Rightarrow 1;$$

$$\textbf{Comfort: } q_{\text{cmf}} = \min_c \sigma(\hat{c};\theta_c),\;\; c \in \{\text{lon-jerk}, \text{lat-acc}, \text{yaw-rate}, \text{accel}, \text{brake}\},$$
$$\theta{=}(4.13,\, 4.89,\, 0.95,\, 2.40,\, 4.05);$$

$$\textbf{Route fid.: } q_{\text{route}} = \min\big(\sigma(\text{xt}_{\text{rms}}; X_0),\, \sigma(\text{hd}_{\text{rms}}; H_0)\big),$$
$$X_0{=}\tfrac{w_{\text{lane}}-w_{\text{veh}}}{2}{=}0.825,\;\; H_0{=}0.09;$$

$$\textbf{Efficiency: } q_{\text{eff}} = \exp\big(-((p-p_H)/\Sigma)^2\big),\;\; \Sigma = \max(0.4\,p_H,\, 4\,\text{m}).$$

Here p_H is the human's progress, and xt_{rms}, $\text{hd}_{\text{rms}}{=}\text{RMS}_t|\text{wrap}(\psi_t^{\text{ego}} - \psi_t^{\text{tangent}})|$ are RMS cross-track and heading deviation from the route centerline.

Every constant comes from outside the planners under comparison. Lane width, the nuPlan/ISO comfort envelope and RSS come from outside the benchmark, and the heading-error $p90$ and the pace referent come from the benchmark's human driver (Table 6, App. G). Kernels fitted to the planners we compare would have made §2's incomparability rate and dimension floor artifacts of our own design. Within an axis, min screens one construct for its worst case and never orders a cross-referent trade. Aggregation *across* axes chooses between constructs, which is the one operation §2 shows cannot be preference-free.

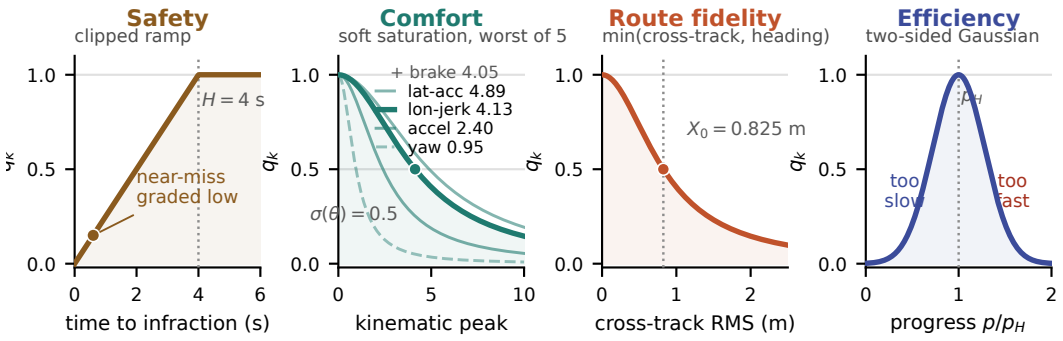


Figure 2: **Anatomy of the four competence axes** (constants as above). Route fidelity takes the worse of the cross-track curve and a heading-vs-tangent term at H_0=0.09 rad, which rewards steering as the road bends.

4.1 Three outputs, three epistemic standings

The released interface returns (g, q, S_p) separately, because the three do not have the same standing. The gate $g \in \{0, 1\}$ answers a legal question and admits no trade-off. The profile $q \in [0, 1]^4$ is the measurement, partial by construction. The scalar is the gated power mean $S_p = g \cdot \left(\sum_k w_k\, q_k^p\right)^{1/p}$ with equal weights and $p \to 0$ by default, and its ordering of the incomparable pairs is declared rather than measured, so we report it with a sensitivity band over (p, w) across which the planner ranking is order-preserving (Table 10, App. I).

The map is the output a scalar-only benchmark cannot give. The scalar judges the incomparable 62–71% (§2), and the map makes that judgement visible. Algorithm 1 has exactly one line of value judgement, which is the line a scalar-only benchmark reports without saying so.

Algorithm 1 Scoring one planner on one scene. Lines 1–3 are measurement; line 4 is the only declared preference.

Require: trajectory τ, scene s with assigned route; horizon H=4 s

1: $g \leftarrow$ AdmissiblePdmRollout(τ, s) ▷ same simulator as PDMS; legal question, no trade-off
2: **if** $g = 0$ **then return** $(0,\ \perp,\ 0)$ ▷ vetoed: never scored, so it cannot be averaged upward
3: $q \leftarrow (q_{\text{saf}},\ q_{\text{cmf}},\ q_{\text{route}},\ q_{\text{eff}})$ as defined above ▷ **the measurement: a vector, never collapsed here**
4: $S_p \leftarrow g \cdot \left(\sum_k w_k q_k^p\right)^{1/p}$ ▷ **declared preference**: equal w, $p \to 0$
5: **return** $(g,\ q,\ S_p)$ ▷ across planners, pairs compare by $\preceq$; incomparable ones stay unordered

5 Validation

The human-versus-blind contrast of §3 showed the profile measuring something a scalar misses, and three questions decide whether that is a readout worth adopting. Does DCP separate planners beyond run-to-run noise, and does that survive the scalarization? Can one game it by maximizing one axis? Does it agree with a referent outside the axes? We answer them on the 13-planner zoo (TransFuser, Latent-TransFuser and EgoStatusMLP at 3 seeds each, the human expert, degenerate controls) plus five SOTA methods (Liao et al., 2025; Zou et al., 2025; Yao et al., 2026; Sun et al., 2026; Li et al., 2025), on the 1098-token `navtest` split (statistics in Table 8, subsets in Table 7).

Separation. Table 2 separates the planner types on route fidelity, efficiency and admissibility, with safety flat at ≈0.98, and the TF-vs-Ego route margin is 4 to 8× the within-type seed spread (Table 9). On the 860 moving scenes the declared scalar orders them human>TF≈LTF>Ego>degenerate in all nine (p, w) settings, both comfort definitions and every seed. On the full split the bottom of that order fails. A straight-line control (0.698) and a constant-velocity one (0.654) pass EgoStatusMLP (0.651), because doing nothing scores well on a near-stationary scene, so we evaluate the scalar on the moving subset.

The Stop row makes the claim easiest to see. A car that does not move really is safer, more comfortable and more on-route than a human expert, and those measurements are correct. Only

Table 2: **DCP profile by planner**, every row from one scoring run (seam-neutral comfort, 1098-token `navtest`, cross-seed mean±std).

Planner	adm. %	Safety	Comfort	Route fid.	Efficiency
Human (expert)	95	0.979	0.918	0.737	1.000
DiffusionDrive (v1)	93	0.980	0.873	0.716	0.808
DiffusionDrive v2	85	0.979	0.538	0.551	0.739
DriveSuprim	86	0.982	0.865	0.653	0.668
SparseDriveV2	92	0.982	0.878	0.669	0.717
WoTE	92	0.980	0.883	0.682	0.807
TransFuser	93	0.980	0.895	0.699±.003	0.830±.006
Latent-TransFuser	93	0.980	0.889	0.694±.007	0.809±.004
EgoStatusMLP (no perc.)	84	0.973	0.883	0.610±.017	0.773±.003
Straight (degenerate)	79	0.974	0.923	0.598	0.981
Constant-velocity	79	0.972	1.000	0.591	0.826
Stop	98	0.990	1.000	0.740	0.233

efficiency separates the two, so the profile reports the pair incomparable in *every* one of the 794 moving scenes where both are admissible, and Stop dominates the human only on 76 near-stationary scenes. Ranking the human above a stationary car is therefore not a measurement but the judgement that getting somewhere counts. Every driving metric makes that judgement, PDMS silently inside its weights and DCP in one declared line, where the scalar reads 0.854 for the human and 0.382 for Stop.

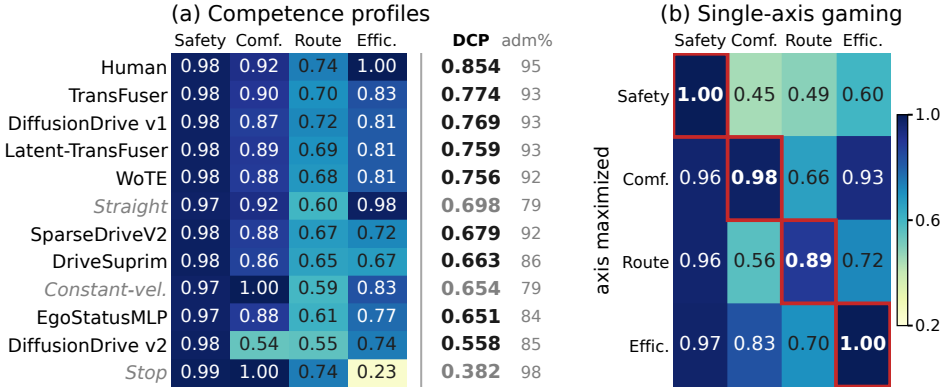


Figure 3: **Profiles and gaming** (seam-neutral comfort). (a) Sorted by DCP; no *aggregate* profile dominates the human's, though Stop does in 7.4% of scenes, all near-stationary. (b) Each axis-maximizing anchor collapses off its own diagonal (Table 11).

Separation is not a total order, and the top of our own leaderboard shows it. DiffusionDrive v1 and TransFuser differ by 0.000, −0.022, +0.018 and −0.022 on the four axes, with mixed signs, so the two leaders are Pareto-incomparable. A scalar must still separate them, and ours does, 0.769 against 0.774, as a declared preference rather than a measurement. This is the pair Fig. 1b sweeps, so the paper's own headline comparison is an instance of invert-or-collapse.

The average TF-vs-Ego route gap of 0.088 is not large, nor should it be, because most `navtest` scenes are easy and there a blind extrapolator really is nearly as good as a planner that sees. Separation concentrates where a planner needs perception. The gap runs from 0.034 on near-straight scenes to 0.134 on high-turn scenes, where the blind model's admissibility also collapses (98% → 67%, Table 13), and Fig. 4 shows what the route axis measures there.

Gaming. A multi-axis readout invites the obvious attack, maximizing one axis and ignoring the rest. Each axis's maximizing anchor collapses on at least one other, so none Pareto-dominates the human (Fig. 3b, Table 11), and added Gaussian noise degrades every axis monotonically (Table 12, App. I).

The subtler attack is that the human-referenced efficiency axis builds the headline in, and it partly succeeds. With a human-free efficiency axis the human-versus-blind rate falls from 83% to 56% [53, 59], below the official PDMS's 71.6%, so the margin over PDMS on that contrast does depend on the human referent. The profile-level separation, carried by the route axis, does not (Table 2). We keep the reference because two-sidedness removes EP's aggression bias, and a one-sided max-progress reference lets the straight-line control dominate the human in 21% of scenes, which is P5 exactly.

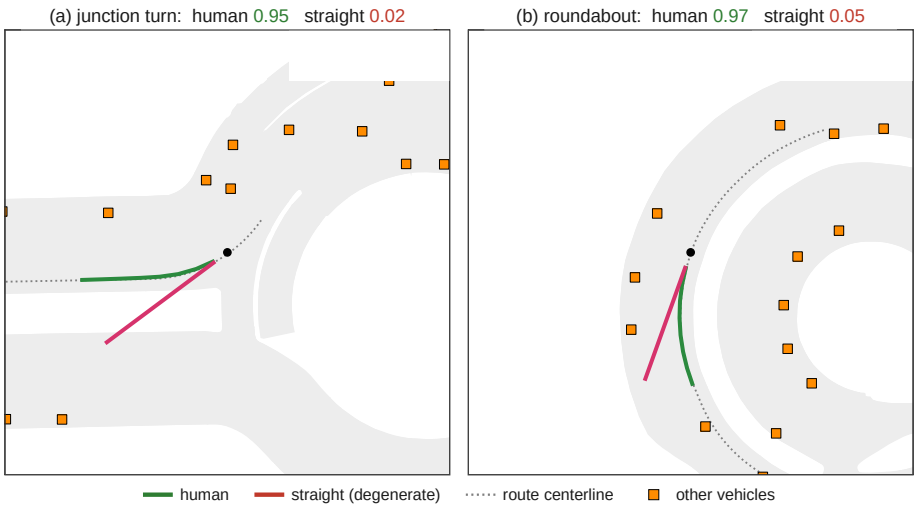


Figure 4: **What the route axis sees**, on the two manoeuvres that separate planners most. The straight-line control scores comfortable (0.96, 0.99) yet inadmissible, smooth and off the route; the human holds the centerline.

External referent. Only the safety axis has one. There, graded safety predicts collisions in a reactive-IDM simulation at AUC 0.887 against 0.59 for the pure-geometry proxy it replaced. Grading buys no accuracy against the binary collision label itself, and we keep it for gradeability (App. K(iv)). The agreement is aggregate rather than per-scene, and on the reactive split (n=644) the gated scalar's Spearman is 0.44, the same as a binary admissibility bit. Comfort has no usable referent (App. E), so the next two subsections test the discriminating axes from outside.

5.1 Evidence of an external Goodhart step

The axes that carry the ranking need a different check, whether DCP sees optimization pressure applied by others for reasons unrelated to our axes. DiffusionDrive v2 (Zou et al., 2025) supplies one, reinforcement learning with "scale-adaptive multiplicative noise . . . to promote broad exploration" and a published PDMS rise from 88.1 (v1) to 91.2.

On the released checkpoints, the only auditable artifacts, the official scorer agrees with DCP on *direction*, and devkit PDMS falls $0.804 \to 0.693$ (CI $[-0.126, -0.095]$). DCP measures route fidelity falling $0.716 \to 0.551$ and comfort $0.873 \to 0.538$, a DCP-scalar gap of $+0.21$ (CI excluding zero under all 9 seed combinations and all three comfort definitions of App. E). The scalar reports the decline as one unattributed number where the profile locates and prices it.

Two facts bound the Goodhart reading. Every figure is a seed mean, and across nine seed combinations the scalar gap moves under 0.001 with axis means moving $\leq$0.013. The published rise also lives in an augmented headline configuration that is not public, so we cannot separate the objective's contribution from the augmentation's, and we flag the Goodhart step rather than assert it (App. K(vi)).

Table 3: **Closed-loop bridge.** Seam-neutral comfort, 1098 tokens, the three Wang et al. (2026) methods with public checkpoints (Yao et al., 2026; Sun et al., 2026; Li et al., 2025). DCP corrects the DriveSuprim/SparseDriveV2 inversion but scores WoTE highest. B2D = closed-loop Bench2Drive (Jia et al., 2024); PDMS_{dk} = official v2 one-stage score; ‡published v1 scores from Wang et al. (2026).

Planner	PDMS‡	B2D‡	adm.%	Safety	Comfort	Route fid.	Efficiency	DCP	PDMS_{dk}
DriveSuprim	93.5	83.02	86	0.982	0.865	0.653	0.668	0.663	0.725
SparseDriveV2	92.0	89.15	92	0.982	0.878	0.669	0.717	0.679	0.734
WoTE	88.3	61.71	92	0.980	0.883	0.682	0.807	0.756	0.800

5.2 Closed-loop bridge

Wang et al. (2026) report PDMS correlating with Bench2Drive (Jia et al., 2024) closed-loop scores across 8 methods (ρ=0.90), but with rank inversions. The most-cited one is on the top pair, where

PDMS ranks DriveSuprim (93.5) above SparseDriveV2 (92.0) while closed-loop puts SparseDriveV2 (89.15) far above DriveSuprim (83.02).

On that pair DCP agrees with closed-loop (Table 3). SparseDriveV2 leads DriveSuprim on every discriminating axis (admissibility 92% against 86%) and on comfort, losing only 0.0007 of safety where nothing discriminates, and the declared scalar orders them 0.679 against 0.663, a gap of $+0.016$ whose paired-bootstrap CI excludes zero. The correction is not weight-repair. By §2.1 the PDMS verdict on this pair is scorer- and weight-fragile, whereas every discriminating DCP axis agrees.

Extend to three planners and the profile stops answering. All three pairs are Pareto-incomparable in aggregate and on 72–74% of jointly-admissible scenes, so a metric returning an order here returns a preference. Ours does. Adding WoTE, worst of the three in closed loop (61.71), the declared scalar puts it *first* at 0.756 against closed-loop's SparseDriveV2>DriveSuprim>WoTE, a Kendall τ of -0.33. The official devkit on the same checkpoints orders the trio WoTE>SparseDriveV2>DriveSuprim (also $\tau = -0.33$, App. J), and only the *published* PDMS reaches $+0.33$, on runs not ours to audit.

Corollary 1 bites here too. The same scalar puts a degenerate straight-line control at 0.698 above *both* SparseDriveV2 (0.679) and DriveSuprim (0.663), and at $n{=}3$ the triple settles only that both behave as §2 says they must.

6 Related work

The ingredients are not new. The joining is, reading a metric's recurring pathologies as evidence about the order type of what it measures (lineage in App. A). Multi-dimensional reporting has precedent (Jia et al., 2024; Liang et al., 2023), and the ML-evaluation literature warns of exactly this failure, because leaderboards commit to one user-utility (Ethayarajh & Jurafsky, 2020), rankings turn on the aggregation rule (Dehghani et al., 2021) and single-number benchmarks encode contested construct choices (Raji et al., 2021).

That conflicting attributes form a partial order, and that dominance ranking is the principled response, founds multi-objective optimization (Deb, 2001) and decision analysis (Figueira et al., 2005), while the non-existence of a neutral scalar aggregator is the subject of social choice (Arrow, 1951; Sen, 1970; Roijers et al., 2013). We instantiate the principle for driving with four grounded axes, two of them newly operationalized (heading-vs-tangent route fidelity, distinct from Li et al. (2024)'s road-boundary check, and two-sided efficiency), and with a reporting framework whose scalar travels with a map of where it judges. Any scoring head is still a map into $(\mathbb{R}, \leq)$ and inherits Proposition 1.

7 Limitations

Two limits are prices of design. Sim-graded safety saturates for competent planners, and human-referenced efficiency under-credits a super-human one. Three more are gaps. DCP is an open-loop proxy blind to compounding closed-loop failure, its safety and gate use PDMS's own simulator, and external validity is thin exactly where discrimination lives, since the only external check anchors on the *saturated* safety axis ($\rho{=}{-}0.11$ without the degenerate anchors, $n{=}446$; two more caveats in App. K).

8 Conclusion

Evaluating driving is a measurement problem, and the field has been scalarizing a partial order. The per-scene poset has certified dimension ≥ 3 in $\sim 91\%$ of scenes under the shipped four-axis partition, so neither one scalar nor two represents it (App. L). Reported at its type, competence keeps what the scalar loses. DCP ranks the human above a blind model in 83% of scenes where the official PDMS manages 71.6%, flags evidence consistent with an external Goodhart step, and still prints a leaderboard number with the map of what it can and cannot arbitrate. Nothing in the argument is specific to driving, and the top-50 models on HELM-Lite and the Open LLM Leaderboard are 86.6–98.9% Pareto-incomparable (App. M).

REPRODUCIBILITY STATEMENT

Every number in this paper resolves to a file in the supplementary material, and the package ships a script that recomputes each checkable one from those files and diffs it against the manuscript. Its `data/` holds the per-scene four-axis arrays in both comfort operationalizations, the same-pool PDMS arrays behind the human-versus-blind comparison, the n=468 arrays behind the in-house per-seed re-rank, and the second scoring run whose disagreement with the shipped one App. E quantifies; its `code/` holds the scoring layer (gate, profile, gated scalar, Pareto relation), the human-free efficiency ablation of App. I, and the fidelity checker. The axis-*extraction* front-end, a PDM rollout to the four axes, additionally requires the NAVSIM devkit and the nuPlan maps, which we cannot redistribute; the arrays above are its output, so every statistic in the paper runs without it. Also included are the order-theoretic outputs (per-scene incomparability and Dilworth widths at $\epsilon \in \{0, 10^{-3}, 10^{-2}\}$ under A and B), the closed-loop pair's per-scene dominance split under both comfort definitions, the compliance-variance decomposition behind P1, and the reweighting families behind §2.1. As an end-to-end check, we assembled the shipped seam-neutral arrays independently on two machines from the released per-planner comfort columns, and they agree bit-for-bit; a second, earlier scoring run of the same planners is also in the artifact, and App. E states exactly where the two disagree rather than leaving it to be discovered. Assumptions and complete proofs for Proposition 1 and Corollary 1 are in App. B; axis calibration constants and their external provenance are in §4 and App. G; split definitions are in App. H.

ETHICS STATEMENT

This work re-analyses an existing public driving dataset and public planner checkpoints; it collects no new driving data and involves no human subjects. A benchmark that ranks driving planners can influence deployment decisions, and we state the limits of that use directly; DCP is an open-loop, single-horizon proxy validated against a reactive simulator on one axis only, and it is a diagnostic complement to closed-loop testing rather than evidence of on-road safety.

AI USE STATEMENT

We used generative AI tools throughout the preparation of this paper, and we disclose that use item by item rather than in summary. *Writing.* They drafted and rewrote sections, which the authors then revised and approved. No section reached its present form without an author rewriting it. *Code.* They acted as coding assistants for parts of the analysis, the re-scoring layer, the figure scripts and the acceptance checkers released with the paper. *Experiment design.* They proposed the partition sweep and null-model battery of App. L, the order-dimension and Dilworth protocols, the closed-loop bridge comparison, and the pre-registered rules by which each was read; the authors decided those rules before the runs. *Interpretation.* They assisted in reading the results, including the diagnosis that our earlier re-scoring layer sampled the simulator five times more coarsely than the devkit (App. K) and that the devkit's reported column is not the one our tooling had been reading. *Retrieval.* They were used to locate and check the cited literature, including the venue and page data in the bibliography. *Formal material.* They assisted with the statements and proofs of App. B and with the numerical verifications reported beside them. We did not use generative AI to produce or impute any reported measurement, and we did not use it to generate data. Every number in this paper comes from the scoring runs described in the text. We executed all AI-assisted code and cross-checked its outputs against independently computed values, we assembled the seam-neutral arrays underlying the headline results on two machines and verified them bit-for-bit, and the released checker recomputes every checkable number in the manuscript from the released files. An author checked every citation against the cited source. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and released artifacts.

REFERENCES

K. J. Arrow. *Social Choice and Individual Values*. Wiley, 1951.

D. Balduzzi, K. Tuyls, J. Perolat, and T. Graepel. Re-evaluating Evaluation. *NeurIPS*, 2018.

H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. *CVPR*, 2020.

H. Caesar, J. Kabzan, K. S. Tan, et al. nuPlan: A Closed-Loop ML-Based Planning Benchmark for Autonomous Vehicles. *CVPR Workshop on Autonomous Driving*, 2021.

W. Cao, M. Hallgarten, T. Li, D. Dauner, X. Gu, C. Wang, Y. Miron, M. Aiello, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, A. Geiger, K. Chitta. Pseudo-Simulation for Autonomous Driving. *Conference on Robot Learning (CoRL)*, 2025.

D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, A. Geiger, K. Chitta. NAVSIM: Data-Driven Non-Reactive Autonomous Vehicle Simulation and Benchmarking. *NeurIPS Datasets & Benchmarks*, 2024.

K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley, 2001.

G. Debreu. Representation of a Preference Ordering by a Numerical Function. In R. M. Thrall, C. H. Coombs, R. L. Davis (eds.), *Decision Processes*, pp. 159–165. Wiley, 1954.

M. Dehghani, Y. Tay, A. A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, D. Metzler, O. Vinyals. The Benchmark Lottery. arXiv:2107.07002, 2021.

R. P. Dilworth. A Decomposition Theorem for Partially Ordered Sets. *Annals of Mathematics*, 51(1):161–166, 1950.

B. Dushnik, E. W. Miller. Partially Ordered Sets. *American Journal of Mathematics*, 63(3):600–610, 1941.

K. Ethayarajh, D. Jurafsky. Utility is in the Eye of the User: A Critique of NLP Leaderboards. *EMNLP*, 2020.

Ö. Evren, E. A. Ok. On the Multi-Utility Representation of Preference Relations. *Journal of Mathematical Economics*, 47(4–5):554–563, 2011.

J. Figueira, S. Greco, M. Ehrgott (eds.). *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, 2005.

M. C. Golumbic. *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, 1980.

Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, H. Li. Planning-oriented Autonomous Driving (UniAD). *CVPR*, 2023.

International Organization for Standardization. ISO 2631-1: Mechanical Vibration and Shock—Evaluation of Human Exposure to Whole-Body Vibration. 1997.

X. Jia, Z. Yang, Q. Li, Z. Zhang, J. Yan. Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-to-End Autonomous Driving. *NeurIPS Datasets & Benchmarks*, 2024.

B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, X. Wang. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. *ICCV*, 2023.

D. H. Krantz, R. D. Luce, P. Suppes, A. Tversky. *Foundations of Measurement*. Academic Press, 1971.

Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, J. M. Alvarez. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? *CVPR*, 2024.

Y. Li, Y. Wang, Y. Liu, J. He, L. Fan, Z. Zhang. End-to-End Driving with Online Trajectory Evaluation via BEV World Model (WoTE). *ICCV*, 2025. arXiv:2504.01941.

P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, et al. Holistic Evaluation of Language Models. *TMLR*, 2023.

B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, X. Wang. DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving. *CVPR*, 2025.

E. A. Ok. Utility Representation of an Incomplete Preference Relation. *Journal of Economic Theory*, 104(2):429–449, 2002.

I. D. Raji, E. M. Bender, A. Paullada, E. Denton, A. Hanna. AI and the Everything in the Whole Wide World Benchmark. *NeurIPS Datasets & Benchmarks*, 2021.

M. Rofin, V. Mikhailov, M. Florinsky, et al. Vote'n'Rank: Revision of Benchmarking with Social Choice Theory. *EACL*, 2023.

D. M. Roijers, P. Vamplew, S. Whiteson, R. Dazeley. A Survey of Multi-Objective Sequential Decision-Making. *JAIR*, 48:67–113, 2013.

A. Sen. *Collective Choice and Social Welfare*. Holden-Day, 1970.

S. Shalev-Shwartz, S. Shammah, A. Shashua. On a Formal Model of Safe and Scalable Self-Driving Cars. arXiv:1708.06374, 2017.

S. S. Stevens. On the Theory of Scales of Measurement. *Science*, 103(2684):677–680, 1946.

M. Strathern. 'Improving Ratings': Audit in the British University System. *European Review*, 5(3):305–321, 1997.

W. Sun, X. Lin, K. Chen, Z. Pei, X. Li, Y. Shi, S. Zheng. SparseDriveV2: Scoring is All You Need for End-to-End Autonomous Driving. arXiv preprint arXiv:2603.29163, 2026.

Y. Wang, A. Jiang, S. Wang, Y. Heng, H. Yang, Y. Chen, H. Sun. Do Open-Loop Metrics Predict Closed-Loop Driving? A Cross-Benchmark Correlation Study of NAVSIM and Bench2Drive. arXiv:2605.00066, 2026.

M. Yannakakis. The Complexity of the Partial Order Dimension Problem. *SIAM Journal on Algebraic and Discrete Methods*, 3(3):351–358, 1982.

W. Yao, Z. Li, S. Lan, Z. Wang, X. Sun, J. M. Alvarez, Z. Wu. DriveSuprim: Towards Precise Trajectory Selection for End-to-End Planning. *AAAI*, 2026. arXiv:2506.06659.

J.-T. Zhai, Z. Feng, J. Du, Y. Mao, J.-J. Liu, Z. Tan, Y. Zhang, X. Ye, J. Wang. Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. arXiv:2305.10430, 2023.

J. Zou, S. Chen, B. Liao, Z. Zheng, Y. Song, L. Zhang, Q. Zhang, W. Liu, X. Wang. DiffusionDriveV2: Reinforcement Learning-Constrained Truncated Diffusion Modeling in End-to-End Autonomous Driving. arXiv preprint arXiv:2512.07745, 2025.

Appendix overview. App. A gives the metric lineage behind §1; App. B, C and D carry the proofs, the order-theoretic measurement and the deployed-pair algebra of §2, and App. E the comfort fork declared there; App. F enumerates the pathologies of §3; App. G states every axis calibration constant and its provenance (§4); App. H and I hold the evaluation splits and full validation tables of §5, App. J the closed-loop bridge of §5.2, App. K the limitations in full (§7), App. L the partition-sensitivity sweep and null controls, and App. M the LLM-leaderboard transfer.

A THE METRIC LINEAGE IN FULL

A bug gets patched; a structural cause survives the patch. The way the field evaluates end-to-end planners has changed three times, and each change fixed the *symptom* of the previous metric.

(i) Open-loop displacement. Benchmarks ranked early E2E planners (Hu et al., 2023; Jiang et al., 2023) by the L_2 displacement and collision rate of a single 3 s open-loop rollout on nuScenes (Caesar et al., 2020). This had no notion of feasibility or interaction, and it credited a planner for reproducing a future that was largely fixed before it acted.

(ii) The shortcut that broke it. Zhai et al. (2023) and Li et al. (2024) showed that a model using only ego status (velocity, acceleration, a discrete driving command, and *no* perception) matches or beats full perception stacks on open-loop L_2, because nuScenes scenes are kinematically predictable. The metric rewarded extrapolating one's own motion, not understanding the scene. Li et al. (2024) additionally proposed a road-adherence metric to catch trajectories that leave the road, answering the exploit by *adding* a measurement rather than re-weighting L_2: the move DCP generalizes, and the check that reappears as its admissibility gate.

(iii) Simulation-aligned scoring. The answer to the shortcut was to make the metric *drive*. NAVSIM (Dauner et al., 2024) replaced L_2 with PDMS, a compliance gate times a weighted mean of quality terms,

$$\text{PDMS}_{\text{v1}} \;=\; \underbrace{\text{NC}\cdot\text{DAC}}_{\text{compliance gate } M} \;\cdot\; \underbrace{\tfrac{5\,\text{EP}+5\,\text{TTC}+2\,\text{C}}{12}}_{\text{weighted quality } W},$$

with a single-frame comfort term C. NAVSIM v2 (Cao et al., 2025) extended it to

$$\text{EPDMS} \;=\; \text{NC}\cdot\text{DAC}\cdot\text{DDC}\cdot\text{TLC} \;\cdot\; \tfrac{5\,\text{EP}+5\,\text{TTC}+2\,\text{LK}+2\,\text{HC}+2\,\text{TFC}}{16},$$

adding two compliance gates (driving direction DDC and traffic light TLC), lane keeping LK, history comfort HC in place of C, a two-frame comfort term TFC computed across consecutive frames, and a human-referenced filter that waives a sub-metric the logged human trajectory itself fails (the human-penalty filter of App. J); the same release added reactive traffic agents and the two-stage pseudo-closed-loop protocol on `navhard`. Each step removed a specific exploit and the successor inherited a new pathology, which is the recurrence charted in §1 of the main text.

Naming in this paper. Every official PDMS number we compute is the v2 devkit's one-stage `score` column on `navtest` (App. J), which is the EPDMS aggregate above without the two-stage protocol and which v2 leaderboards report as EPDMS; we write PDMS_{dk} where the distinction matters (Tables 1 and 3). EPDMS in Table 1 and §3 denotes the reactive two-stage protocol. Published PDMS values quoted for other planners (§2.1, §5.1, Table 3) are v1 scores; where the two appear side by side only the ordering is compared, and App. J accounts for the difference in level.

Social-choice-grounded aggregation offers declared alternatives to the implicit weighted mean: Nash averaging (Balduzzi et al., 2018) and voting-rule leaderboards (Rofin et al., 2023) choose a principled completion, where DCP instead reports what any completion decides; the two are complementary, and either could serve as DCP's declared scalar.

B PROOFS

Throughout, a scalar metric factors through the attribute vector, $m = f \circ q$ with $f : [0,1]^4 \to \mathbb{R}$; Pareto-monotonicity is monotonicity of f in the dominance order on q. Without this restriction two Pareto-monotone scalars could disagree on a pair with equal attribute vectors, and the "exactly P_{inc}" clauses of Proposition 1(a,b) would fail on such pairs; with it, equal-vector pairs are ties for every m considered. Here $\mathcal{A}_s$ is the admissible set of scene s, $q : \mathcal{A}_s \to [0,1]^4$ the competence attribute map of Definition 1, and $\preceq$ the induced componentwise partial order. A scalar metric is any $m : \mathcal{A}_s \to \mathbb{R}$; it is *Pareto-monotone* if $\tau \prec \tau' \Rightarrow m(\tau) < m(\tau')$. Write P_{inc} for the set of incomparable pairs and $I_s = |P_{\text{inc}}|/\binom{N}{2}$.

Lemma 1 (No faithful scalarization)**.** *If $(\mathcal{A}_s, \preceq)$ has an incomparable pair, no m is faithful, because $(\mathbb{R}, \leq)$ is total, so m asserts $<$, $=$, or $>$ on that pair, a relation $\preceq$ omits.*

Proof. Let $\{\tau, \tau'\}$ be incomparable, so neither $q(\tau) \leq q(\tau')$ nor $q(\tau') \leq q(\tau)$ holds componentwise. The order $(\mathbb{R}, \leq)$ is total, so exactly one of $m(\tau) < m(\tau')$, $m(\tau) = m(\tau')$, $m(\tau) > m(\tau')$ holds, and $\preceq_m$ therefore asserts a relation between τ and τ'. Each of the three assertions is absent from $\preceq$: the two strict ones because neither vector dominates, the equality because a faithful representation would require $q(\tau) = q(\tau')$, which contradicts incomparability. Hence m is not a homomorphism of $(\mathcal{A}_s, \preceq)$ into $(\mathbb{R}, \leq)$. □

Proof of Proposition 1. **(a)** Apply the argument of Lemma 1 pairwise. Any m whatsoever misrepresents every pair in P_{inc}, giving the bound $\geq I_s$ with *no* monotonicity assumed. Now let m be Pareto-monotone and let $\{\tau, \tau'\}$ be comparable, say $\tau \prec \tau'$. Monotonicity gives $m(\tau) < m(\tau')$, which is exactly the relation $\preceq$ asserts, so no comparable pair is misrepresented and the misrepresented set is exactly P_{inc}. Partition it by the value m asserts. Pairs receiving a strict order are *inverted* in the sense that some equally legitimate scalar orders them the other way (by (b)), pairs receiving equality are *collapsed*. These are exhaustive and disjoint, so $\text{inv}(m) + \text{col}(m) = |P_{\text{inc}}|$. Dropping monotonicity can only add misrepresented comparable pairs, never remove incomparable ones.

(b) Let $d = q(\tau') - q(\tau) \in \mathbb{R}^4$ and consider the positive-linear family $m_w(\cdot) = \langle w, q(\cdot)\rangle$ with $w > 0$ componentwise. Two members disagree on the pair iff $\langle w, d\rangle$ and $\langle w', d\rangle$ have opposite signs for some admissible $w, w' > 0$. If $d \geq 0$ or $d \leq 0$ (the comparable case) then $\langle w, d\rangle$ has a fixed sign for every $w > 0$, so no disagreement is possible. If d has a strictly positive entry $d_i > 0$ and a strictly negative entry $d_j < 0$ (the incomparable case), take w concentrating mass on coordinate i and w' concentrating mass on coordinate j: the two inner products have opposite signs. Since the positive-linear family is contained in the legitimate (Pareto-monotone) family, the disagreement set of the latter contains and, by (a), is contained in P_{inc}.

(c) An antichain of $(\mathcal{A}_s, \preceq)$ is a set of pairwise-incomparable elements. Let w_s be the width, i.e. the size of a maximum antichain (Dilworth, 1950). That antichain contributes $\binom{w_s}{2}$ incomparable pairs, all of them in P_{inc}, so $I_s = |P_{\text{inc}}|/\binom{N}{2} \geq \binom{w_s}{2}/\binom{N}{2}$. The bound uses only the order relation, with no axis scaling, no weights and no choice of m. □

Proof of Corollary 1. Let $\{\tau, \tau'\} \in P_{\text{inc}}$. By totality of $(\mathbb{R}, \leq)$, m asserts a strict order or a tie. In the strict case, part (b) supplies a second legitimate scalar asserting the opposite strict order, so the assertion is a commitment to one member of a family whose members disagree, a preference rather than a measurement. In the tie case, m identifies two elements that $\preceq$ distinguishes on at least two coordinates with opposite signs. There is no third option, and no choice of m avoids both. □

Two consequences used in §2 follow immediately. First, a Pareto-monotone m has per-pair faithfulness at most $1 - I_s$, attained iff it represents every comparable pair, so the observed ceiling is a property of the poset rather than of any fitting procedure. Second, because (b) characterizes disagreement by the sign pattern of d alone, the question "can re-weighting fix this pair?" is decidable without sampling weights, which is what §2.1 exploits.

C ORDER-THEORETIC MEASUREMENT IN DETAIL

Deciding dimension ≤ 2. A poset has Dushnik–Miller dimension ≤ 2 iff its incomparability graph is transitively orientable (Dushnik & Miller, 1941; Golumbic, 1980). This is decidable exactly, so per scene we certify $\dim \geq 3$ rather than estimate it; we do not claim the exact value beyond 3, since deciding $\dim \leq k$ for $k \geq 3$ is NP-complete (Yannakakis, 1982). Dimension 1 means the poset is a chain, i.e. a scalar represents it faithfully; that occurs in no `navtest` scene.

Stability under re-operationalization. Re-defining all four axes across 256 combinations of alternative kernels leaves the incomparable rate between 56% and 87%, with 34% of pairs called incomparable by every one of the 256. Across ten operationalizations the $\dim \geq 3$ share stays between 79.8% and 92.1%. Because Pareto dominance depends only on the per-axis order, the rate is additionally invariant under *any* strictly-monotone reparameterization of an axis, since rescaling a kernel cannot change it at all.

Tolerance. Calling exact ties "comparable" is the conservative direction, and the conclusion survives it. Among the eleven ranked planners the incomparable fraction is 71.9% (A) / 71.5% (B) at $\epsilon{=}0$ and 55.2% (A) / 51.9% (B) at $\epsilon{=}10^{-2}$. That tolerance is about half the ≤ 0.019 seed-to-seed spread in route fidelity, so it does not by itself absorb run-to-run noise; it does show the majority is not an artifact of testing exact equality. The weight-free Dilworth floor over the same planners runs from 35.7% at $\epsilon{=}0$ to 16.9% at $\epsilon{=}10^{-2}$.

Two correlation summaries, and why they are not the argument. Correlation is a weaker instrument than the order-theoretic question, since two axes can be uncorrelated and still admit a scalar that orders every pair correctly. Reported for comparison only, the correlation spectrum's participation ratio is 2.83 (A) to 3.76 (B), while inverting the independence null $P(\text{incomparable}) = 1 - 2^{1-k}$ on the observed rate gives about 2.8 ordering-independent axes; weak coupling suppresses ordering-independence faster than it shrinks the linear spectrum. Neither is a fitted dimension. Safety and route fidelity are orthogonal under both definitions ($r{=}0.02$).

The scalar ceiling is attained. Proposition 1(a) caps per-pair faithfulness at $1 - I_s$. A dense positive-weight power-mean family saturates that ceiling exactly at 0.380, and a monotone gradient-boosted regressor, given the axes as features and fitted to reproduce the relation, reaches 0.379, falling short by tying some comparable pairs because it is only weakly monotone. Neither represents a single incomparable pair, which is the content of the proposition rather than a shortcoming of the fit.

D THE DEPLOYED-PAIR ALGEBRA

§2.1 states that PDMS's verdict on DriveSuprim/SparseDriveV2 is scorer- and weight-fragile. This appendix gives the quantities behind that, which turn the claim into an exact interval per scorer variant.

The differences themselves, under the official scorer. On the 1060 of 1098 common tokens whose stored sub-metric columns reconstruct the reported `score` exactly, the gated-mean difference $\delta_k = \mu_k(\text{SparseDriveV2}) - \mu_k(\text{DriveSuprim})$ per quality term of the official aggregation ($5/5/2/2/2$ over EP, TTC, LK, HC, TFC) is

$$\delta_{\text{EP}} = +0.0421, \quad \delta_{\text{TTC}} = -0.0250, \quad \delta_{\text{LK}} = -0.0415, \quad \delta_{\text{HC}} = -0.0255, \quad \delta_{\text{TFC}} = +0.0920.$$

The reachable gap over all positive weightings, the convex hull $[-0.0415, +0.0920]$, *spans zero*. The official weights recombine the δ_k to $+0.0085$ on the clean tokens, and the stored scores put SparseDriveV2 ahead by $+0.0092$ over all 1098: the closed-loop side, but well under DCP's $+0.016$. Our coarser $8@0.5$ rescoring lands the *other* side (-0.018; all five in-house δ_k negative: $-0.0029/-0.0187/-0.0178/-0.0187/0.0000$), and the published scores call it the other way again. The positive hull mass sits on ego progress and two-frame comfort; the compliance-adjacent terms all favour DriveSuprim. Closed loop separates this pair by 6.1 points; every scalar reading of it we can audit sits within ± 0.02 and changes sign with scorer implementation or weighting. Under the reactive policy (scored at $8@0.5$ only) the in-house construction gives $[-0.0373, -0.0180]$.

What EPDMS's extra term buys. Under our in-house proxy the two-frame comfort column is uncomputable on a single-frame split and returns a constant ($\delta_{\text{TFC}} = 0.0000$), so there weight placed on it only drags the gap toward zero. The official pipeline *does* compute it, via scene adjacency, and on this pair it is the largest single term ($\delta_{\text{TFC}} = +0.0920$, toward SparseDriveV2). The added term does not stabilize the verdict, it hands the sign to whichever side the newest, sparsest-coverage column happens to favour.

Where the near-inertness comes from. Under the official scorer, ego progress is the only continuously graded quality term (717 and 614 distinct values across the two planners); TTC, LK, HC and TFC are per-scene binary. Our coarser $8@0.5$ scoring silenced TTC and HC entirely (App. J), which is where an earlier draft's "seven of 14 weight units on zero-variance terms" reading came from; the devkit version of that statement is that seven units sit on binary-or-constant terms whose grading is a coin per scene.

Non-linear and two-stage variants. The convex-hull argument covers weighted means exactly, so no sweep is needed there. For non-linear aggregators we sampled, under the $8@0.5$ scoring, power means at $p \in \{0.25, 0.5, 1, 2, 4\}$, 20,000 weightings each, under both traffic policies, return flip fraction 0.0000 throughout (closest approach at p=0.25, gap -3.5×10^{-7}), as does the $50/50$ two-stage EPDMS proxy; under the devkit deltas the linear analysis above already characterizes the flip region exactly, so these sweeps are historical.

The one structure that does flip it. Dissolving the multiplicative gate into the sum (scoring all nine sub-metrics as one weighted average, which is no longer PDMS) changes the sign pattern. TLC and EP become positive ($+0.0009$ and $+0.0159$), the reachable interval becomes $[-0.0222, +0.0114]$, and 8.7% of the simplex flips the pair (6.7% under reactive agents). That is the exception that proves the point, because recovering this comparison takes a change of structure rather than a change of weights.

E THE COMFORT A/B FORK

Comfort is the one axis admitting two defensible operationalizations, and they move the headline numbers, so we carry both throughout and adopt one on a test external to the axes.

Definition A (as-scored) seeds the acceleration series with the ego's current velocity, so its first sample mixes a genuine start transient with a first-pose seeding mismatch. *Definition B (seam-neutral)* drops that sample and scores the interior series. *Definition C (simulated rollout)* recomputes comfort on the simulator's own states.

Adjudicating by convergence with NAVSIM's comfort sub-score would be circular, because that sub-score is computed on the simulated rollout, and definition C is a graded version of the same quantity on the same states, so C's high convergent AUC is not evidence in its favour. (The shipped definition B converges with that sub-score at AUC 0.71. The sub-score flags only 8 of the $1{,}464$ trajectories as uncomfortable, so it cannot discriminate either way.) We therefore adjudicate on a *dose–response* test instead: inject a known, externally imposed noise dose into the trajectory and ask which definition resolves adjacent doses most cleanly (Table 4). Definition B separates adjacent doses at a minimum effect size of 0.979 standard deviations, against 0.904 for A and 0.388 for C. All three are monotone in the dose; B wins on resolution, which is the property a comfort axis needs.

Table 4: Comfort dose–response (183 scenes, $1{,}464$ trajectories): mean comfort at each injected noise level (Gaussian, std), and the minimum effect size between adjacent levels. We report convergent AUC but it does not adjudicate, for the circularity noted above.

	Noise (m)						
Definition	0.0	0.15	0.3	0.5	0.8	min. effect	conv. AUC
A as-scored (ego-seeded)	0.910	0.655	0.386	0.230	0.115	0.904	0.920
B seam-neutral **(shipped)**	0.888	0.691	0.435	0.261	0.142	**0.979**	0.713
C simulated rollout	0.958	0.913	0.855	0.759	0.615	0.388	0.992

What the comfort fork does and does not move. Every number in this paper comes from one scoring run, the one the release ships. A second run of the same planners exists in the artifact. The two agree bit-for-bit on safety and comfort, but they disagree on route fidelity on nearly every scene (max $|\Delta|$ up to 0.27) and flip admissibility on 0–20 scenes per planner. The head-to-head rate reads 84.2% there against 83.4% here.

The second scoring run, diagnosed. We diagnosed the route-fidelity disagreement rather than only disclosing it. Rebuilding the axis from the shipped source function, fed the re-run's *own* stored trajectories, reproduces the *shipped* column to floating point on all 7 of 13 planners with deterministic trajectories, and it reproduces the re-run's column on none, under the shipped pose grid and three alternatives. The six sensor-based planners differ by at most $3e-02$, an order below the anomaly and consistent with GPU inference non-determinism between runs. The shipped run is therefore the code-consistent one, and the re-run's route column is not consistent with its own stored trajectories under any variant of the function that produced it. Attributing the 0.8-point head-to-head spread to the comfort fork would in any case have been wrong. Swapping only the comfort definition, every other axis held fixed, moves the rate by $+0.000$ inside the shipped run and $+0.003$ inside the other, so this statistic is essentially blind to the fork. The fork does bite where we report it (incomparability, 62% against 71%), which is why we carry both there.

The choice is consequential and we report both everywhere. Under A the incomparable fraction is 62% and the comfort–efficiency correlation is $r{=}0.83$ (past our own redundancy flag, the seam artifact surfacing as correlation), while under B they are 71% and $r{=}0.21$. The participation ratio moves from 2.83 (A) to 3.76 (B). What does *not* move is the claim the paper rests on, since the order-dimension floor is 91% under A and 90.5% under B. The released axis arrays ship in both operationalizations; for provenance the raw ego-seeded variant A remains the default of the extraction front-end, with a seam-neutral switch.

F THE PATHOLOGY ENUMERATION

Our internal enumeration runs P1–P7. The main text reports the four most directly implied by the partial-order structure; Table 5 collects them with their splits. We list the remaining three here for completeness; DCP's design addresses two of them rather than using them as evidence for it.

- **P3 non-reactivity.** Standard `navtest` replays background agents that never react to ego. Re-scoring under a reactive IDM policy, 5.6% of trajectories PDMS rates perfect collide once agents react. EPDMS (Cao et al., 2025) addresses this directly, which is why we do not use it as evidence, and every result in this paper that could be affected is also reported under reactive agents.
- **P4 single-scalar inadequacy.** One number cannot say *where* a planner is weak. This is not a separate pathology so much as the thesis of the paper, made precise by Proposition 1.
- **P6 binary cliff.** PDMS's gates are near-binary and multiplicative, so a marginal single infraction yields the same 0 as a high-speed crash. DCP's graded safety axis de-cliffs this by construction; App. K reports what that grading does and does not buy.

What the P1 runs are. The variance decomposition uses 45 NAVSIM evaluation logs we had on hand, all of them runs of our own planner family rather than a sample of the field, 37 from one architecture's sweep, 6 validation runs, and 2 controls. Three cover full `navtest` (12,082 tokens) and give 82.4, 84.1 and 84.1%. The other 42 are 267-token subsets with median 77% and range 63–96%, the spread being what a 267-scene sample of a variance ratio looks like. We quote the full-run figure in the main text because it is the one that is not sample-limited, and report the subset spread here rather than averaging the two populations into a single number. The claim P1 makes, that most of PDMS's variance is spent on legality, does not depend on which of the two is used.

P7 is the per-scene consequence of P2 and P6 rather than an independent root cause, and we label it as such. Figure 6 draws P1 and P2 and Figure 5 draws P7, both quantified in Table 5. Figure 7 carries a result the main text has no room for. Among 5585 high-PDMS, error-free trajectories, PDMS is near-uncorrelated with lane discipline, naturalness and anticipation ($|\rho| \leq 0.02$), 73% of that set are worse than the human on at least two of six quality measures and 50% on at least three, and PDMS ranks the human above its own aggressive maximizer in 0.1% of 868 scenes against 47.2% for the competence axes.

Table 5: Four PDMS pathologies, quantified. Full `navtest` unless noted; P7 under the official devkit scorer on the 1098-token split (App. J).

Pathology	Statistic	Value
P1 compliance dominates	gate share of $\mathrm{Var}(\text{PDMS})$, 3 full runs	82–84%
	same, 42 subset runs: median (range)	77% (63–96%)
P2 saturation	compliant scenes; mean PDMS there	95.7%; 0.934
	PDMS std: all / compliant	0.188 / 0.076 (2.5×)
P5 aggression bias	EP at $0.5/1.0/1.5\times$ human extent	0.904/0.951/0.966
	EP awarded the human	0.836
	scenes: aggressive anchor $>$ human	99.8%
P7 unrankability	human–TF separation < 0.01 / < 0.05	46.3% / 68.2%
	per-scene winner = / contradicts aggregate	48.8–63.6% / 19.7–28.8%
	scenes tying human, TF and blind exactly	13.0% (ours: 47.2%)

G AXIS CALIBRATION AND EXTERNAL PROVENANCE

Figure 2 in the main text draws the four calibration curves.

Table 6 records where each constant comes from. The point of the table is negative, since no entry is fitted to the planners being separated, so the incomparability rate of Measurement 1 and the dimension floor of Measurement 2 cannot be artifacts of tuning our own kernels.

Two axes are new operationalizations rather than re-scalings of existing sub-metrics. Route fidelity pairs cross-track error with a heading-vs-tangent term, $\mathrm{hd}_{\mathrm{rms}} = \mathrm{RMS}_t|\mathrm{wrap}(\psi_t^{\mathrm{ego}} - \psi_t^{\mathrm{tangent}})|$; the alternative $\sum|\Delta\psi|$ used elsewhere rewards never steering (§4); together with the lane gate this term removes the straight-line exploit. Efficiency is two-sided about p_H, so it penalizes both loitering

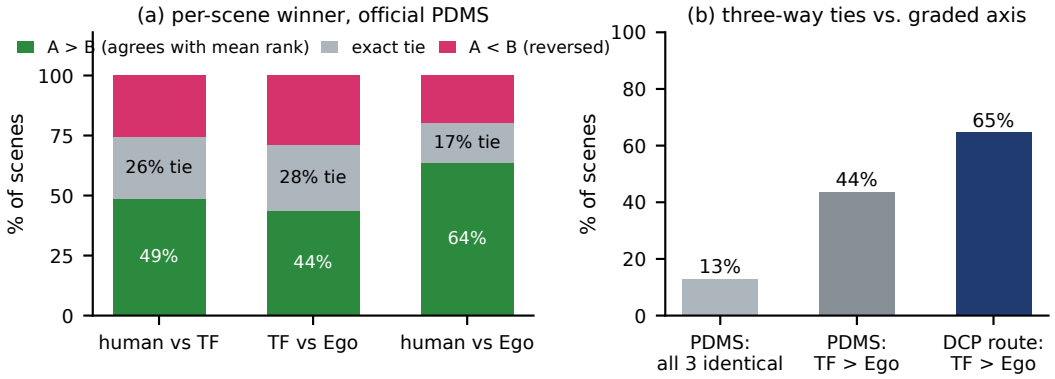


Figure 5: **P7, per-scene unrankability** (official devkit scorer). (a) The mean order survives per scene in only $48.8/43.7/63.6\%$ and reverses in 19.7–28.8%. (b) Exact three-way ties, 13.0%; the graded route axis recovers 65% TF>Ego.

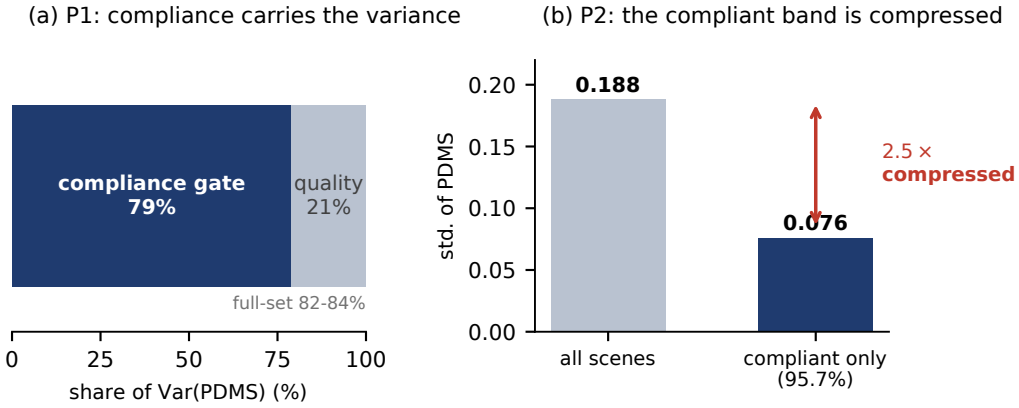


Figure 6: **P1/P2.** The compliance gate carries $\sim 80\%$ of $\mathrm{Var}(\mathrm{PDMS})$, and within the compliant band the spread is $2.5\times$ compressed against the ceiling.

and progress-at-any-cost; PDMS's EP normalizes by the most-aggressive proposal in the scene and is therefore one-sided (P5 in App. F).

H Evaluation splits

Every experiment runs on `navtest`; Table 7 records which subset and its size. Splits differ because the underlying artifacts do. The order-theoretic analysis needs a candidate pool with many trajectories per scene, whereas the planner comparison needs one trajectory per planner on a common token set.

[†]Each named subset is the token set on which the corresponding arrays in the artifact are defined; every array ships with its token list, so each subset is reconstructible from the files it names.

I Full validation tables

Does the headline survive a human-free efficiency axis? Efficiency is referenced to the human, who therefore scores 1.000 on it by construction, so the sharpest objection to the head-to-head result is that the axis makes it true. We test that directly. Holding the statistic fixed (seed-averaged blind model, gated geometric mean, strict per-scene comparison, scene bootstrap $B{=}2000$) and replacing efficiency with a human-independent reference centred on the best progress any committed planner achieves in that scene, the rate falls from 83.4% $[81.2, 85.6]$ to 55.7% $[52.8, 58.8]$. A one-sided progress ratio, PDMS's own EP form, gives 56.9% $[54.1, 59.8]$; restricting the reference to a seven-

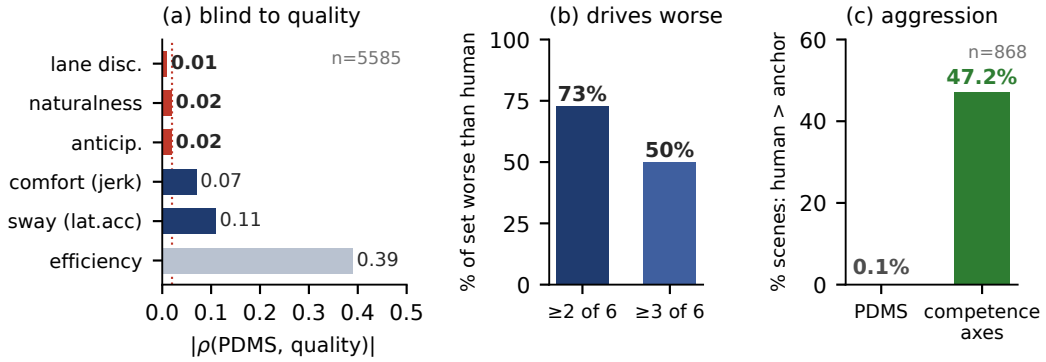


Figure 7: **High score, drives poorly.** Among high-PDMS, error-free trajectories PDMS (our scorer; App. J) is near-uncorrelated with naturalness, lane discipline and anticipation ($|\rho| \leq 0.02$), and its aggressive maximizer beats the human.

Table 6: Provenance of every calibration constant. None is estimated from the planner set under comparison.

Axis / gate	Constant	Source
Admissibility	τ_{lane}=1.75 m (half lane)	nuPlan map lane geometry
Safety	horizon H=4 s, time-to-infraction	RSS (Shalev-Shwartz et al., 2017); nuPlan (Caesar et al., 2021)
Comfort	θ=(4.13, 4.89, 0.95, 2.40, 4.05) over lon-jerk, lat-acc, yaw-rate, accel, brake	nuPlan comfort envelope / ISO 2631 (ISO, 1997)
Route fidelity	$X_0=(w_{\text{lane}} - w_{\text{veh}})/2$=0.825 m	lane and vehicle width
Route fidelity	H_0=0.09 rad	measured human heading-error $p90$
Efficiency	p_H, $\Sigma=\max(0.4p_H, 4\,\text{m})$	the human expert's own progress

planner roster (the human, one seed each of TransFuser, Latent-TransFuser and EgoStatusMLP, and the three controls) moves neither by more than 0.2 points.

The human-referenced axis therefore carries roughly half the margin. The scorer correction changed what the ablation licenses. Under our coarser scorer PDMS reaches 37.6% [34.9, 40.4] on the same scenes, and the human-free DCP beats that by 19 points; under the official reported column PDMS reaches 71.6% [68.9, 74.2], which the human-free variant does *not* exceed. The margin over the official PDMS on this contrast therefore depends on the human referent, and the main text says so; the profile-level separation (route axis) does not. We report the human-referenced axis as the default because the two-sided form is what removes EP's aggression bias (P5).

The human against a stationary car. Under the shipped seam-neutral comfort, Stop scores exactly 1.000 on comfort in all 1098 scenes, so the human can never weakly dominate it on all four axes and therefore never strictly dominates it at all. Over the 1030 jointly-admissible scenes the split is 0.0% human-dominates, 7.4% Stop-dominates (76 scenes), 92.6% incomparable. Restricted to the 794 moving scenes (human progress $\geq$ 3 m) it is 0%/0%/100%: every one of the 76 dominations is a near-stationary scene, where matching the human's pace costs nothing, and among the 236 near-stationary jointly-admissible scenes that is 32.2%. Substituting a human-free efficiency axis leaves all of these figures identical, so the human-referenced axis is not the cause.

Under the as-scored comfort (A) the same comparison reads 43.0% human-dominates against 2.5%, on 1050 jointly-admissible scenes, because A's ego-seeded first sample denies Stop a perfect comfort score. That reversal is an artifact of the seam rather than a judgement about driving, which is why the shipped definition is the one the main text discusses.

Table 7: **Evaluation splits** (all `navtest`): which experiment runs on which subset, and its size.

Use	Split	Size
Order-theoretic structure (§2)	sim-safety candidate pool	484 sc., 10,386 cand.
Planner profiles (Tab. 2, 9)	common learned, SOTA, controls	1098
In-house per-seed re-rank (App. I)	lidar re-rank set†	468
Current SOTA (Tab. 2, 3)	DiffusionDrive v1/v2, DriveSuprim, SparseDriveV2, WoTE	1098 each
Convergent comfort / safety AUC	comfort / safety arrays	1,464 / 776
Reactive-sim Spearman (§5)	reactive candidate set	644
Gaming, monotonicity (Tab. 11, 12)	sim-graded anchor set	432 of 500
Closed-loop case study (§5.2)	DriveSuprim, SparseDriveV2, WoTE	1098 each

Per-seed head-to-head rates (in-house scorer). Both aggregates score the seed-averaged EgoStatusMLP. Per seed, our coarser PDMS ranks the human above it in 31.0, 38.0 and 34.0% of scenes against DCP's 78.6, 85.6 and 79.1%, so the lowest DCP seed still exceeds the highest PDMS seed under that scorer; on the lidar re-rank set it averages 0.590 PDMS for EgoStatusMLP against the human's 0.701, ranking the human above it in only 31% of scenes. Averaging the three per-seed rates is a different statistic from scoring the seed-averaged model and does not reproduce the aggregate.

Table 8 is the complete battery behind §5; the main text quotes from it rather than reprinting it.

Table 8: Validation summary in full: every test with its statistic. A/B are the two comfort operationalizations; brackets are 95% CIs.

Property	Quantity	Value
Partial order	incomparable admissible pairs (A/B)	62/71%
	scenes with order dim ≥ 3 / $= 1$	$\sim$91% / 0%
	participation ratio (A/B)	2.83/3.76
Ranked planners (11)	incomparable, ϵ=0 (A/B)	71.9/71.5%
	incomparable, ϵ=10^{-2} (A/B)	55.2/51.9%
	Dilworth floor (weight-free), ϵ=0 $\rightarrow$ 10^{-2}	35.7 $\rightarrow$ 16.9%
Construct stability	incomparable over 256 operationalizations	56–87%
	unanimous across all 256	34%
Orthogonality	max axis $\lvert r\rvert$, comfort–efficiency (A/B)	0.83/0.21
Per-axis validity	human vs. stop, efficiency axis	1.000 / 0.233
Discrimination	route gap TF$-$Ego vs. seed std	0.088 / $\leq$0.019 (4–8$\times$)
Head-to-head (1098)	human $>$ blind, DCP, shipped run	83.4% [81.2, 85.6]
	human $>$ blind, official PDMS	71.6% [68.9, 74.2]
	human $>$ blind, PDMS, in-house scorer	37.6% [34.9, 40.4]
	same, human-free efficiency (two-sided / ratio)	55.7 / 56.9%
	effect of the comfort fork alone, within a run	+0.000 / +0.003
	same statistic on the second scoring run	84.2%
Robustness	scalarizations preserving order (moving-scene)	9/9
Gaming	single-axis anchor Pareto-dominates human	0%
	same, human-free efficiency	$\leq$9%
Convergent (safety)	AUC vs. reactive collision (geometry-only)	0.887 [0.85, 0.92] (0.59)
External (Goodhart)	published PDMS, DiffusionDrive v1$\rightarrow$v2	88.1 $\rightarrow$ 91.2
	DCP scalar gap, v1$-$v2	+0.21 [+0.20, +0.23]
Closed loop (n=2)	DCP gap, SparseDriveV2$-$DriveSuprim	+0.016 [+0.003, +0.029]
Closed loop (n=3)	Kendall τ: DCP / devkit PDMS / published	$-$0.33 / $-$0.33 / +0.33

Table 9 gives every real planner at every seed, the source of the by-type means in Table 2: within-architecture seed variation on the discriminating axes is at most 0.019 in route fidelity, so the between-architecture gaps are signal. Table 10 shows the declared scalar is order-preserving across power means and weight schemes (9/9). Table 11 is the gaming matrix: each single-axis maximizer is high on its own axis and collapses elsewhere, and none Pareto-dominates the human; only under the human-free efficiency variant of App. I can a single-axis anchor dominate the human at all, in at most 9% of scenes (the $\leq$9% row of Table 8). Its anchors are per-scene optima rather than a fixed policy, so it and Table 12 run on the 432-scene anchor set of Table 7 rather than the 1098. Table 12 shows every axis degrading monotonically as Gaussian noise is added to the human path,

with axis-appropriate slopes (comfort and route steep, efficiency flat because zero-mean noise barely changes net progress). Table 13 stratifies the discrimination by scene difficulty, and Figure 3 draws the profile and the gaming matrix together, with the declared scalar per planner in its right-hand column, including the 0.698 of the straight-line control discussed in §5.2.

Table 9: **Per-seed profiles, nine learned planners** (1098 scenes). Seed variation on route fidelity, the discriminating axis, is ≤ 0.019.

Planner (seed)	adm.%	Safety	Comfort	Route fid.	Eff.
TransFuser-0	93	0.980	0.897	0.699	0.837
TransFuser-1	93	0.979	0.894	0.695	0.824
TransFuser-2	93	0.981	0.895	0.702	0.829
LTF-0	93	0.979	0.886	0.686	0.815
LTF-1	93	0.981	0.890	0.693	0.806
LTF-2	93	0.980	0.890	0.703	0.805
EgoStatusMLP-0	86	0.972	0.885	0.631	0.769
EgoStatusMLP-1	83	0.974	0.875	0.589	0.778
EgoStatusMLP-2	84	0.974	0.887	0.611	0.772

Table 10: **Declared-scalar ranking robustness** (gated power mean S_p; moving-scene subset, 860 of 1098 scenes). The order holds in all 9 $(p, \text{weights})$ settings; five representative ones are shown. On the full split the straight control inflates on near-stationary scenes.

Weights	p	Human	TF	Ego	Straight
equal	1.0	0.857	0.816	0.724	0.658
equal	0.5	0.851	0.808	0.715	0.651
equal	0	0.844	0.800	0.703	0.643
safety-heavy	1.0	0.855	0.818	0.726	0.655
route-heavy	0	0.801	0.752	0.654	0.603

Table 11: **Gaming matrix** (production sim-graded safety): each row is the anchor maximizing one axis. Every anchor is high on its diagonal and blocked off it: "dom. human" is 0% throughout, and the safety-maximizer is inadmissible in 37% of scenes.

Maximized axis	Safety	Comfort	Route fid.	Efficiency	adm.%	dom. human
Safety	**1.000**	0.446	0.487	0.597	63	0%
Comfort	0.961	**0.978**	0.659	0.934	85	0%
Route fid.	0.961	0.560	**0.885**	0.718	99	0%
Efficiency	0.966	0.828	0.697	**1.000**	89	0%

J CLOSED-LOOP BRIDGE IN DETAIL

Table 3 gives the three methods from Wang et al. (2026) with public `navtest` checkpoints, with both the published scores and the official-devkit rescoring of our runs. Two readings must be kept apart. On the DriveSuprim/SparseDriveV2 pair, DCP agrees with closed-loop and the *published* PDMS does not; the official rescoring of our runs also lands on the closed-loop side, by a near-tie margin ($+0.0092$) that §2.1 shows is scorer- and weight-fragile. Across all three methods, neither scalar matches the closed-loop order.

Per scene, SparseDriveV2 strictly dominates DriveSuprim in 23% (comfort A) / 20% (B) of scenes against 8% / 10% the reverse, with 69% incomparable under both. Restricting to the 903 jointly-admissible scenes gives $16/10/74\%$ (A) and $15/12/74\%$ (B) for dominate / dominated / incomparable. The aggregate Pareto verdict on the pair is technically incomparability, but only because SparseDriveV2 is lower on safety by 0.0007 (0.9818 against 0.9825) on the axis that discriminates nothing; every discriminating axis favours SparseDriveV2.

The declared scalar inherits the pathology the paper describes, and we state the worst case rather than leave it implicit: DriveSuprim's DCP scalar of 0.663 is low enough that the degenerate straight-line control out-scores it, strictly dominating it in 24% of jointly-admissible scenes against under 1% the other way. This is an argument for reading the profile rather than the scalar, not evidence that the

Table 12: **Monotonicity** (production sim-graded safety): profile vs. Gaussian noise added to the human path (std, metres). Every axis degrades monotonically, comfort and route steeply, efficiency and safety gently. Same 432-scene set as Table 11.

Noise (m)	adm.%	Safety	Comfort	Route fid.	Eff.
0.00	94	0.967	0.908	0.791	1.000
0.25	93	0.952	0.489	0.440	0.999
0.50	90	0.935	0.230	0.221	0.996
1.00	40	0.874	0.086	0.072	0.986
2.00	2	0.861	0.026	0.018	0.943

Table 13: **Difficulty stratification.** DCP's TF−Ego separation concentrates on hard (turning) scenes, where the blind model's admissibility also collapses.

Stratum	route gap (TF−Ego)	TF>Ego %	TF adm.%	Ego adm.%
Straight (low turn)	0.034	44	99	98
Turn (high)	**0.134**	**72**	85	67
Interactive (low margin)	0.087	65	94	88

straight-line control drives well, since it is inadmissible in 21% of scenes and bottom of the route axis.

The published-versus-recomputed gap deserves its own statement, and its cause is now identified. The devkit constructs its PDM simulator and scorer on a proposal sampling of 40 poses at 0.1 s and passes that sampling to the scoring call. Our pipeline used a single constant of 8 poses at 0.5 s for both the trajectory representation and the scoring call, a five-times coarser simulation grid. Scoring identical trajectories on the identical metric cache at 8@0.5 reproduces our stored PDMS exactly (max per-scene difference 0 over all 1098 tokens), while scoring at the devkit's 40@0.1 differs on 733–794 of them. The sampling constant is the entire discrepancy, and the devkit itself warned of it once per token (an observation of length 9 where 41 was expected).

The mechanism is under-detection. A violation between two coarse samples is never seen, so the coarse grid over-scores compliance; at 2 Hz the time-to-collision and history-comfort sub-scores never registered a violation at all. The correction is not a uniform shift (it moves the human by -0.155, TransFuser by -0.068 and EgoStatusMLP by $+0.028$), which is why the uncorrected implementation misordered published pairs. Ranked by disagreement mass the residual falls first on ego progress (0.099), then lane keeping (0.053), with drivable-area compliance near the bottom (0.031). Of DCP's components only the safety axis touches the simulator; the admissibility gate, route fidelity, comfort and efficiency are computed from trajectory geometry alone and are unaffected.

The constant is left uncorrected in the shipped pipeline for this submission, since correcting it silently redefines every simulator-derived number. Wherever PDMS itself is the claim, this paper quotes the devkit CLI's reported column (`score`: components after the human-penalty filter, two-frame comfort injected by the scene aggregator), not the `pdm_score` column those CLIs explicitly exclude, which is pre-filter and internally inconsistent once the filter fires.

The split itself is sound, since the human trajectory scores 0.945 on the full official `navtest` token list and 0.943 on our 1098-token split under the reported column, so the split is not a harder population. Our reproductions of the public checkpoints do score below their published full-`navtest` numbers under either column. Part of that gap is the formula, since the published numbers are v1 PDMS (App. A), and the same checkpoints score 2–6 points lower under the v2 aggregate in their own papers (SparseDriveV2 92.0 → 90.1 under the v2 release that applies the human-penalty filter, as our column does; DriveSuprim 93.5 → 87.1 under the earlier release without it; Sun et al., 2026; Yao et al., 2026); the remainder is on our inference side, under investigation, and every cross-planner comparison in this paper is within our own matched pipeline. Under devkit scoring the trio orders WoTE (0.800) > SparseDriveV2 (0.734) > DriveSuprim (0.725), Kendall $\tau = -0.33$ against closed-loop, matching our uncorrected scorer's $\tau = -0.33$.

K LIMITATIONS IN FULL

§7 states these compactly; here is each in full.

(i) DCP is an open-loop proxy. On the three methods from Wang et al. (2026) with public `navtest` checkpoints we run a closed-loop case study. DCP corrects the most-cited PDMS↔closed-loop inversion but does not order all three planners as closed-loop does, and *neither does* PDMS recomputed on the same split (§5.2). An open-loop profile cannot see compounding closed-loop failure; a full correlation needs closed-loop scores the field has not released for enough public checkpoints.

(ii) With a faithful sim-graded safety, the safety axis saturates for competent planners. It catches degenerate behaviour (validated by the gaming matrix, the monotonicity sweep and the collision AUC) but does not rank good planners; discrimination lives in route fidelity, comfort and efficiency, plus the admissibility gate.

(iii) Human-referenced efficiency treats the skilled human as the competent pace, so a super-human planner is under-credited on that axis. The alternative is worse, since a human-independent max-progress reference lets the degenerate straight-line control dominate the human in 21% of scenes.

(iv) De-cliffing the collision test buys no collision-prediction accuracy, since the tie-averaged AUC is 0.887 either way (n=776). The "0.92 binary check" cited in earlier drafts is not a competing predictor, since it is the collision label scored against itself under a different traffic policy, so its AUC of 1.0 is automatic. The case for grading is gradeability, not accuracy.

(v) DCP's safety and gate use the same PDM simulator as PDMS; we improve the readout, not the simulation.

(vi) The Goodhart evidence carries two caveats. The v2 trajectories we score come from the released *no-augmentation* configuration, the only v2 artifact in our possession; the published 91.2 PDMS is the authors' own number for their headline configuration, so the score side of the published rise rests on a variant we cannot audit, and we cannot attribute the measured collapse to the reinforcement-learning objective alone. On the released artifacts themselves the official devkit *agrees* with DCP about the direction ($0.804 \to 0.693$, CI $[-0.126, -0.095]$; the same sign under our coarser sampling, -0.024, so the conclusion does not depend on the scorer defect of App. J). The devkit's comfort term registers the collapse at $0.99 \to 0.82$, about half the $0.873 \to 0.538$ DCP measures and further diluted by the weighted mean, so the scalar under-prices rather than misses the damage; our earlier statement that the comfort collapse could not register in PDMS was an artifact of the 2 Hz reimplementation and is withdrawn.

(vii) The competence construct carries content validity throughout, but external and convergent validity is thin exactly where discrimination lives. It is anchored only on the *saturated* safety axis, at aggregate reactive-collision AUC 0.887. Even there, the degenerate and random anchors carry the per-scene agreement with the reactive simulator. Excluding them, the Spearman turns slightly negative, mean ρ=-0.11 over n=446 scenes, and -0.14 on the hardest near-anchors (n=444), against $+0.44$ $[0.41, 0.47]$ when all anchors are included, the same 0.44 a binary admissibility bit achieves. Graded safety therefore adds ranking information over a binary collision bit in aggregate but not scene by scene. The three axes that carry *all* cross-planner discrimination (route fidelity, comfort, efficiency) have content and face validity (Fig. 4) but no independent external referent. Establishing convergent validity against independent human or closed-loop judgments remains future work. Collapsing back to a scalar removes none of these limits.

L PARTITION SENSITIVITY AND NULL CONTROLS

Measurements 1 and 2 are computed on a chosen partition of the raw channels into four axes. The 256-operationalization sweep of App. C varies kernels *within* that partition; this appendix varies the partition itself, then asks whether the observed structure exceeds what decoupled channels would produce. Everything here runs on the 484-scene candidate pool under ego-seeded comfort A, the extraction default (App. E). Everything reuses the exact $\dim \leq 2$ decision procedure; every headline was cross-checked by a second, independent decision path on a 20-scene subsample, with zero incomparability or dimension disagreements and zero 2-realizer certificate failures. The artifact retains the superseded null constructions alongside the final ones, with pre-registration stamps, so the number of nulls tried is visible.

The sweep. Merging uses the paper's own within-axis rule ($\min$ of members); the four shipped axes were first verified to *be* the $\min$ of their raw channels to 10^{-16} (and provably not the mean or product), so merging and splitting are exact inverses of the shipped aggregation. Table 14 reports every 2+2 pairing, every merge-one-pair k=3 variant, and channel-level splits. Three facts:

1. *The incomparability rate is partition-relative and monotone in arity*, from 26.6–42.4% at k=2 through 62.1% at the shipped k=4 to 86.3% at k=8; a reviewer who expects this is right, and the main text states the rate as a property of the four-axis operationalization, not of driving per se.
2. *The dimension floor is not monotone: it peaks at the shipped partition* (90.9%, against 66.3% at k=8 and 52.9% at k=10). The mechanism is internal to order dimension. As a poset approaches a total antichain its incomparability graph approaches K_n, which is transitively orientable, so $\dim$ falls back to 2; this was confirmed on synthetic independent axes with no connection to the pool. More axes buy incomparability and *spend* dimension.
3. *Merging safety is nearly free* (84.7–89.5% $\dim \geq 3$ retained at k=3) *while merging any pair from* {*comfort, route, efficiency*} *collapses the floor* (to 6.0–8.5%), so the discriminating structure lives in three axes. Indeed the safety axis is constant across all candidates in 246 of 484 scenes (50.8%), the P2 saturation of §3 made exact.

Table 14: Partition sweep on the candidate pool (484 scenes, 10,386 candidates); merges use the paper's own $\min$ rule. All values in %. †one safety channel is constant, so k=10 is effectively k=9.

partition	k	inc. (ϵ=0)	inc. (ϵ=10^{-2})	$\dim$ =2	$\dim \geq 3$	Dilworth floor
k=1 (any scalar)	1	0.0	0.0	0.0	0.0	2.0
{saf,rte}{cmf,eff}	2	42.4	34.0	100.0	0.0	8.6
{saf,cmf}{rte,eff}	2	30.8	23.8	100.0	0.0	6.8
{saf,eff}{cmf,rte}	2	26.6	8.9	99.4	0.0	5.9
merge saf+rte	3	57.9	40.2	10.5	89.5	19.5
merge saf+cmf	3	58.8	40.8	10.7	89.3	19.8
merge saf+eff	3	58.2	42.1	15.3	84.7	19.4
merge cmf+rte	3	33.7	13.3	94.0	6.0	7.5
merge cmf+eff	3	47.7	38.8	91.5	8.5	10.4
merge rte+eff	3	36.7	29.4	93.4	6.6	8.1
shipped k=4	4	62.1	44.8	9.1	90.9	21.9
split comfort (k=8)	8	86.3	60.9	33.7	66.3	50.8
full split (k=10)†	10	89.7	68.5	47.1	52.9	58.9

Table 15: The same instrument on public LLM leaderboards (downloaded 2026-09-04; provenance in the artifact). All values in %. *dimension certified on the top-200 by mean score; incomparability and Dilworth are full-set.

source	set	n	attrs	inc. (ϵ=0)	inc. (ϵ_{rel}=10^{-2})	dim	Dilworth floor
Open LLM Lb. v2	all	4576	6	54.1	51.2	$\geq$3*	2.9
Open LLM Lb. v2	top50	50	6	98.9	97.2	$\geq$3	80.8
Open LLM Lb. v2	top20	20	6	99.5	97.9	2	100.0
HELM Lite 1.13	all	90	10	69.9	66.8	$\geq$3	16.6
HELM Lite 1.13	top50	50	10	86.6	82.9	$\geq$3	33.1
HELM Lite 1.13	top20	20	10	98.4	96.3	2	80.5
HELM Classic 0.4	all	29	16	78.6	74.9	$\geq$3	33.5
HELM Classic 0.4	top50	29	16	78.6	74.9	$\geq$3	33.5
HELM Classic 0.4	top20	20	16	86.3	82.6	$\geq$3	47.9
HELM Classic (dense)	all	65	10	70.4	65.9	$\geq$3	14.4
HELM Classic (dense)	top50	50	10	78.3	73.6	$\geq$3	24.5
HELM Classic (dense)	top20	20	10	94.7	90.5	2	71.6

Null controls. We use two nulls, a within-scene *independent* permutation of each channel (preserves every marginal, destroys all coupling) and a *physically constrained* block null that permutes each axis's channels together (preserving within-axis joint distributions, destroying only cross-axis association, which is what the claim is about). Comparisons are made at *matched incomparability*; matching on arity instead biases the two statistics in opposite directions (the null sits several points more incomparable, which suppresses its $\dim \geq 3$ and inflates its Dilworth floor), a confound that two early arity-matched runs exhibited and that reversed under matching – both are retained in the artifact as superseded. Findings:

1. *The headline rate is conservative.* Observed incomparability sits *below* the independent null at every partition (-14.8 points at $k{=}4$), since decoupling the same marginals produces more incomparability rather than less, so the rate was not installed by the choice of axes.
2. *The disagreement is clustered, not diffuse.* At matched density, incomparability-graph transitivity, mean local clustering and max-antichain pair share all exceed both nulls (twelve of twelve gaps positive, $+0.012$ to $+0.155$), and the Dilworth floor exceeds the physical null at every headline partition ($+9.2$/$+8.9$/$+7.0$ points at $k{=}4/8/10$), so real trade-offs concentrate into large Pareto fronts.
3. *The* $\dim \geq 3$ *share sits modestly below the matched null* (-3.7 points at $k{=}4$, the comparator there interpolated between measured null points whose two independent constructions agree to 0.031 mean absolute difference; -11.8 at $k{=}8$ with measured local comparators). A synthetic probe shows this is the *predicted* signature of clustered incomparability, since a union-of-cliques graph family reproduces the clustering surpluses and the orientability (hence lower-dimension) tendency at once, while cross-edge noise destroys orientability first. We therefore do *not* offer $\dim \geq 3$ as evidence that the partition is non-arbitrary; that evidence is items 1–2. The $\dim \geq 3$ statement itself needs no null, because it is an exact property of the measured poset, certified per scene.

Three limits remain. The incomparability-matched independent null carries the target density but not the pool's exact marginals; the block null still assumes axes could be freely recombined; and the mechanism probe uses 24-vertex synthetic graphs, establishing the mechanism qualitatively rather than fitting the pool.

M THE ORDER AUDIT BEYOND DRIVING

The instrument of §2 is not driving-specific, since any leaderboard that averages per-attribute scores into one number can be audited the same way. Table 15 applies it to public LLM leaderboards, with the same exact $\dim \leq 2$ certification and the same two-path verification. The transfer is not marginal. Among the top-50 models by mean score, 86.6% of HELM-Lite pairs and 98.9% of Open-LLM-Leaderboard pairs are Pareto-incomparable – more than the driving pool's 62.1% – and $\dim \geq 3$ is certified on most model sets. The top-20 sets instead certify $\dim = 2$ exactly, because they are near-total antichains (Dilworth floor 100% on the Open-LLM top-20), the K_n regime of App. L, which is a stronger statement against scalar ranking than $\dim \geq 3$ (almost no model dominates another) but is not the $\dim \geq 3$ statement and is not reported as one. The Fig. 1b construction also transfers. Two adjacent, currently deployed models on HELM Lite are Pareto-incomparable, and a weighted mean flips their order at $w{=}0.41$, closed form agreeing with a grid sweep. Three caveats apply. Leaderboard rows are evaluated configurations rather than de-duplicated models; the largest poset's dimension is certified on its top-200; and the sparse HELM-Classic release is reported under both missing-value policies, which agree qualitatively.