RETHINKDRIVE: SUPERVISION TARGETS FOR LIGHTWEIGHT TRAJECTORY REFINEMENT IN END-TO-END DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Candidate-based driving planners must select one trajectory from many plausible alternatives. A candidate's driving score can guide this choice, but does not establish its value as a training target. We present RethinkDrive, a lightweight refinement procedure and controlled study of this distinction. With the planner and candidate pool frozen, we compare metric-derived targets with targets anchored to recorded human trajectories, which are used only in training. We evaluate learned rankings before and after combining their scores with the planner's. On NAVSIM, reference supervision improves a diffusion planner, mainly by making successive plans more temporally consistent, with some loss of progress and drivable-area compliance. The gain persists with disjoint training and validation logs and without score fusion; two validation-selected rules do not reproduce it. Fusion even reverses the comparison between a hard reference target and a validation-selected soft target on this planner. On a vocabulary-based planner, it reduces large differences between learned rankings to similar final scores, without a resolved improvement over the baseline. These findings show that supervision quality must be assessed through both the learned ranking and its use in the deployed selection rule.

1 INTRODUCTION

A high-scoring trajectory is not necessarily a useful training target for a trajectory selector. The target first shapes a learned ranking; deployment may then combine this ranking with an existing planner's scores. We show that these two steps can change which target appears most effective. RethinkDrive studies them separately while keeping the planner and its candidates fixed, and uses reference supervision to refine the final choice.

End-to-end driving learns actions from observations and navigation commands (Codevilla et al., 2018; Wu et al., 2022). Its representations range from camera–LiDAR fusion (Chitta et al., 2023) and bird's-eye-view features (Li et al., 2022; Hu et al., 2022) to joint scene models (Casas et al., 2021; Hu et al., 2023; Jiang et al., 2023) and latent world models (Li et al., 2025b). Candidate-based planners separate trajectory generation from selection (Liao et al., 2025; Xing et al., 2025). A useful candidate improves driving only when selected. Refining this last choice can exploit an existing planner without retraining perception or generation.

Driving scores and demonstrations provide different supervision. A score summarizes collision avoidance, progress and comfort, whereas distance to a demonstration measures agreement with one recorded behavior. Planners already learn from both (Li et al., 2024; Yao et al., 2026). The unresolved issue here is how converting these signals into discrete or distributed targets affects a learned selector, and how score fusion changes the resulting comparison. Using human trajectories alone is therefore insufficient to explain a method's benefit.

This distinction matters when many candidates share the maximum score: taking the first maximum assigns a label by candidate order, although their trajectories differ. Ties do not imply unsafe behavior, and a demonstration need not be optimal. We therefore measure label ambiguity, compare target rules under shared training conditions, and test the learned ranking separately from the fused decision (Figure 1).

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.

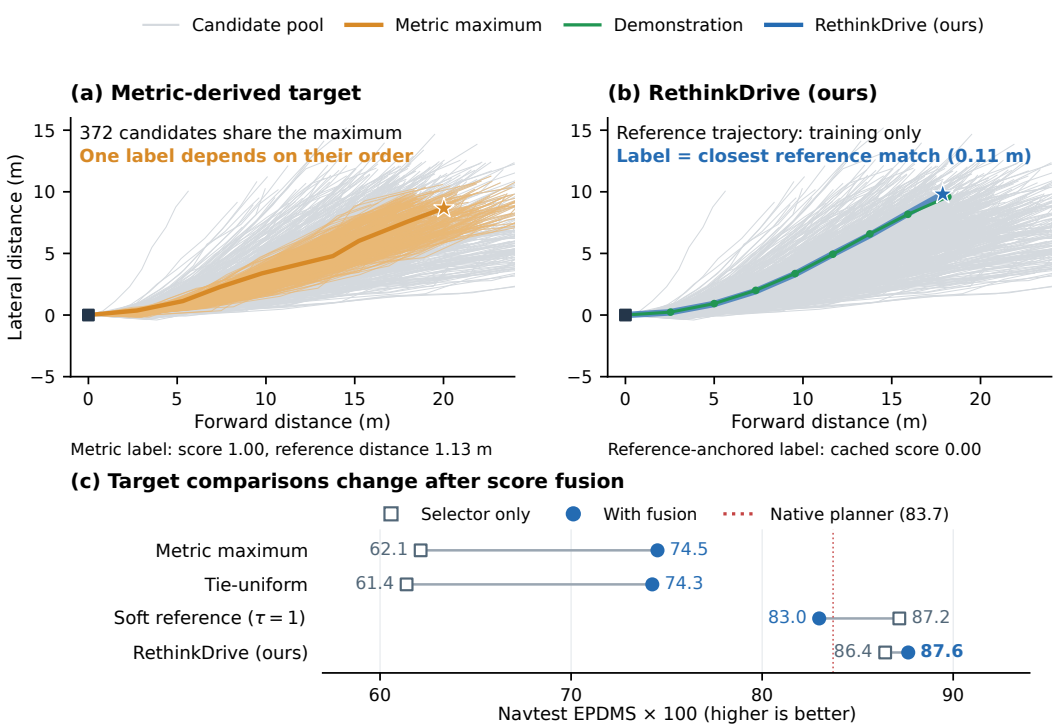

Figure 1: **Target construction and score fusion are distinct parts of selection.** (a–b) In one validation scene, 372 of 800 candidates share the maximum cached score; the nearest-reference label instead scores 0. Neither label score describes learned performance. (c) DiffusionDriveV2 navtest extended Predictive Driver Model Score (EPDMS) $\times 100$, averaged over three seeds. Each connector compares the same learned selector before and after fusion. Reference supervision improves native selection; fusion reverses the hard–soft ordering (Table 4). References are training-only.

RethinkDrive audits score saturation and ties, constructs a minimum-displacement reference target, and trains a lightweight attention selector. Its output corrects the frozen planner's score. Here *refinement* revises the choice among existing trajectories; it does not generate or alter waypoints. This design makes target rules directly comparable.

On DiffusionDriveV2 (DDv2), metric-maximum supervision scores 9.2 percentage points (pp) below native selection, while RethinkDrive gains 3.9 pp on navtest and 1.6 pp on navhard. The navtest gain survives log-disjoint retraining and removing fusion, and two training-free rules do not match it. These benefits concern temporal comfort, with progress and compliance costs. On DriveSuprim, fusion instead brings different learned rankings near the baseline. Thus target design and the execution rule must be assessed together.

Contributions.

- **Distinguishing label score from training utility.** We quantify tied maxima in two candidate pools and show that the higher-scoring metric labels train a weaker DDv2 selector. Tie resolution and partial metric alignment recover only part of this deficit (Sections 3 and 5.3).
- **Lightweight refinement with controlled evidence.** A reference target and 0.52 M-parameter selector improve DDv2 without retraining its planner. Log-disjoint retraining, fusion removal and rule comparisons establish the scope of this gain (Sections 4 and 5).
- **A distinction between learned and executed rankings.** Same-checkpoint comparisons show that fusion reverses hard–soft rankings on DDv2 and compresses target differences on DriveSuprim, despite substantial changes in selected trajectories (Section 5.5).

2 Related Work

Learning driving behavior. Conditional imitation (Codevilla et al., 2018), privileged supervision (Chen et al., 2020) and learning from other agents (Chen & Krähenbühl, 2022) use demonstrations differently. ChauffeurNet augments imitation with perturbed trajectories (Bansal et al., 2019). PlanT uses object inputs (Renz et al., 2023); ThinkTwice refines plans with scalable decoders (Jia et al., 2023); DriveTransformer unifies task, sensor and temporal attention (Jia et al., 2025). We keep the scene encoder fixed.

Candidate generation and selection. Trajectory sets and anchors represent multimodal behavior in forecasting (Phan-Minh et al., 2020; Chai et al., 2020). For planning, DiffusionDrive and DDv2 use truncated diffusion (Liao et al., 2025; Zou et al., 2025), GoalFlow uses flow matching (Xing et al., 2025), and Diffusion Planner supports closed-loop planning (Zheng et al., 2025). VADv2 models a planning vocabulary (Jiang et al., 2026); SparseDrive rescores collision risk (Sun et al., 2025), while SparseDriveV2 factorizes paths and velocities (Sun et al., 2026). Li et al. (2026) train scorers on augmented samples from frozen planners. Our comparisons instead isolate how target construction and score fusion affect the choice within a fixed candidate pool.

Supervision and evaluation. Hydra-MDP combines human and metric supervision (Li et al., 2024; 2025a); DriveSuprim uses coarse-to-fine selection and self-distillation (Yao et al., 2026). Learned planning costs (Zeng et al., 2019; Huang et al., 2024b;a), safety reasoning (Kim et al., 2026) and preference learning (Shang et al., 2025) intervene beyond label construction. Distillation (Hinton et al., 2015) and ranking objectives (Burges et al., 2005; Xia et al., 2008; Rafailov et al., 2023) provide established learning tools. Our evidence concerns target construction and its interaction with execution, using these tools. Soft reference is a distance-target control, not a Hydra-MDP reproduction; metric-maximum labels do not represent all metric distillation. Driving-system biases (Jaeger et al., 2023; Dauner et al., 2023) and imitation-learning distribution shift (Ross et al., 2011) motivate separating offline agreement from driving performance. Appendix F details the distinctions.

3 Supervision Ambiguity

For scene s, a frozen planner proposes K trajectories a_i with predicted scores f_i. Separately, cached Predictive Driver Model scores (PDMS) combine no at-fault collision (NC), drivable-area compliance (DAC), ego progress (EP), time-to-collision (TTC) and comfort. The extended score (EPDMS) adds further criteria; Appendix A defines both. Saturation means the scene maximum reaches the score ceiling. Ties are candidates within 10^{-6} of that maximum.

Table 1: **Driving scores often leave the target unresolved.** Median tied-set size is conditional on a tie. The aligned rows recompute all 800 candidates using single-stage EPDMS without two-frame extended comfort (EC). Pool precision and score definitions differ; the rows do not rank planners.

Candidate source / score	Split (scenes)	K	Saturated	Tied	Median
DiffusionDriveV2 / PDMS	Validation (5,278)	800	88.2%	85.3%	134
DiffusionDriveV2 / PDMS	Training (87,392)	800	88.1%	85.3%	145
Same pool / aligned EPDMS	Validation (5,278)	800	**82.5%**	**84.7%**	98
Same pool / aligned EPDMS	navtest (12,146)	800	74.4%	77.4%	81
DriveSuprim / EPDMS	navtrain (103,288)	8,192	91.8%	90.5%	223

The first exact maximum is the lowest tolerance-tied index in every validation scene. In 60 audited sets, only 0.8% of candidate pairs are within 0.1 m; mean endpoint spread is 1.25 m. DriveSuprim reproduces the diagnosis on another vocabulary, although half-precision caching can increase ties.

Aligned scoring removes pose padding and changes 50.2% of validation maximizing indices, yet ties persist on both splits. Cached-score mismatch alone therefore cannot explain the ambiguity. This does not explain the training deficit; Section 5.3 tests tie resolution separately from reference supervision.

4 METHOD

RethinkDrive refines a frozen planner's choice (Figure 2). Offline auditing and reference targets train only the selector; inference combines its output with the planner's scores.

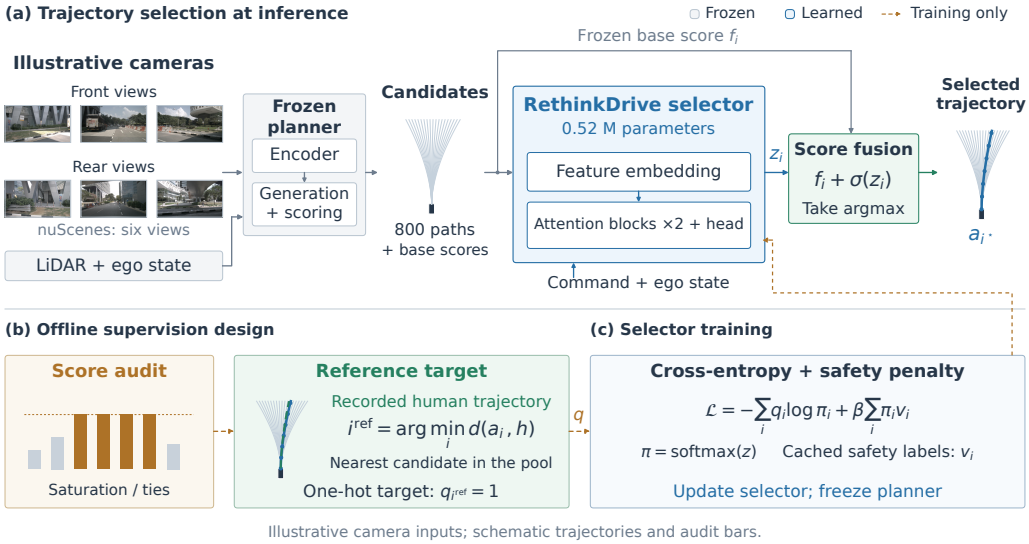

Figure 2: **RethinkDrive architecture and supervision.** (a) The frozen planner supplies candidates and scores; the learned selector adds a sigmoid correction. (b) Score auditing motivates reference targets; the arrow denotes a design rationale. (c) Cross-entropy and an expected safety penalty train only the selector; dashed arrows are training-only. Context is constant in training and recorded at evaluation (Section 4.2). Six nuScenes views (Caesar et al., 2020) illustrate surround observations, not the evaluated sensor configuration. Curves and audit bars are schematic; experiments use NAVSIM.

4.1 SCORE AUDIT AND REFERENCE-ANCHORED TARGET

Before training, the audit measures score saturation, tie frequency and tied-set size. Shared maxima make first-index labels depend on candidate order, motivating an alternative target. The audit is a dataset analysis; it neither learns parameters nor runs at inference.

For the reference target, let $h_s(t)$ denote the recorded human position at step t, with $T = 8$. We measure time-aligned displacement and choose the nearest candidate:

$$d_i(s) = \frac{1}{T} \sum_{t=1}^{T} \|a_i(t) - h_s(t)\|_2, \qquad i^{\text{ref}}(s) = \arg\min_{i \le K} d_i(s). \tag{1}$$

The minimum is nearly unique: within 10^{-6} m, ties occur in one of 5,278 validation scenes and two of 4,370 sampled training scenes. A unique minimum is invariant to candidate permutation. These measurements support practical insensitivity to order, rather than an unconditional guarantee for the rare ties. Distance is a supervision signal for recorded behavior, not a safety measure.

4.2 FROZEN CANDIDATES AND SELECTOR INPUTS

Our DiffusionDriveV2 (Zou et al., 2025) configuration uses ResNet-34 (He et al., 2016) and supplies 800 candidates: 200 samples from 20 anchors and 10 noise draws, plus three rescaled groups. Each has eight waypoints at 0.5 s intervals. Selector inputs are 16 geometry features, six predicted scores, command and ego state; the reserved visual slot is zero-filled. Absent training signal files make command and ego state constant in all 87,392 training scenes. Validation and primary test instead supply recorded commands and $[v_x, v_y, 0, 0]$. A separate control repeats training and checkpoint selection with constant context throughout. Navtest means shift by at most 0.27 pp and retain their ordering (Appendix G); the primary results keep their original protocol.

4.3 FUSED LIGHTWEIGHT SELECTOR

Candidate features comprise geometry g_i and base predictions b_i: coarse score, NC, EP, DAC, TTC and comfort. Two multilayer perceptrons embed $[g_i, b_i]$ and scene context to width 128. Two attention blocks (Vaswani et al., 2017) use cross-attention to context, self-attention across candidates and a 128–256–128 feed-forward network. Each has four heads, dropout 0.1, residual connections and layer normalization. A linear head produces logit z_i. The measured module contains 0.519 M parameters.

Writing $\pi_i = \mathrm{softmax}(z)_i$ and $q_i^{\mathrm{ref}} = \mathbf{1}[i = i^{\mathrm{ref}}]$, the loss is

$$\mathcal{L}(s) = -\sum_i q_i \log \pi_i + \beta \sum_i \pi_i v_i + \gamma c_{\mathrm{rank}} \mathcal{L}_{\mathrm{rank}}(s), \tag{2}$$

where $v_i = (1 - \mathrm{NC}_i) + (1 - \mathrm{DAC}_i) + (1 - \mathrm{TTC}_i)$ uses cached simulation labels, $\beta = 0.3$ and $\gamma = 0.2$. RethinkDrive and all four main targets set $c_{\mathrm{rank}} = 0$. Cross-entropy teaches the target distribution, while the expected safety penalty discourages probability on cached violations. Gradients pass through softmax probabilities, without differentiating a discrete argmax. Sharing this penalty across the main targets keeps their constraint supervision fixed.

At inference, the selector corrects the frozen coarse score f_i through sigmoid fusion:

$$i^\star(s) = \arg\max_i [f_i + \lambda \sigma(z_i)], \qquad \lambda = 1. \tag{3}$$

Here σ denotes the logistic sigmoid. Only the selected index changes; candidate waypoints, generator and base scoring heads remain unchanged. The *coarse baseline* sets $\lambda = 0$. The *native baseline* uses the same checkpoint's own coarse-to-top-32 fine-scoring path. Comparison with native selection measures the gain over the existing planner. Executing $\arg\max_i z_i$ separately measures the learned ranking without fusion; this distinction matters because training uses softmax and execution uses sigmoid.

Information and computation at execution. The planner's predicted sub-scores are available inputs; cached simulation sub-scores are training labels only. This distinction prevents access to future evaluation outcomes at inference. Recorded human trajectories are likewise used only by the offline target builder. The selector's batch-1 forward pass takes 0.56 ms on an RTX PRO 6000, excluding candidate generation and feature extraction (Table 18). Its cost therefore measures the added module, rather than the full driving pipeline.

4.4 MATCHED TARGET CONTROLS AND COMPONENT VARIANTS

Let m_i be cached PDMS and $\mathcal{T}(s) = \{i : m_i \geq \max_j m_j - 10^{-6}\}$. The four main targets differ only in their probability assignment:

$$\begin{aligned} q_i^{\mathrm{Metric}} &= \mathbf{1}[i = \text{first} \arg\max_j m_j], \quad & q_i^{\mathrm{Uniform}} &= \frac{\mathbf{1}[i \in \mathcal{T}(s)]}{|\mathcal{T}(s)|}, \\ q_i^{\mathrm{ref}} &= \mathbf{1}[i = \arg\min_j d_j], & q_i^{\mathrm{Soft}} &= \frac{\exp(-d_i/\tau)}{\sum_j \exp(-d_j/\tau)}. \end{aligned} \tag{4}$$

Uniform supervision removes the first-index preference among near-maximal candidates. Soft supervision weights reference displacement; the hard target is its $\tau \to 0$ limit when the minimum is unique. Validation selects $\tau = 1$ m from $\{0.1, 0.3, 1\}$ m. This is a diffuse target: median mass on the nearest candidate is 0.4%, with effective support about 470 of 800 candidates. Test sensitivity retains the validation choice.

A tie-resolution control uses $i^{\mathrm{tie}} = \arg\min_{i \in \mathcal{T}(s)} d_i$: the reference chooses only among metric maxima. A support ablation restricts target construction to the first 80 unperturbed candidates; inference still scores all 800. Within this subset, a length variant keeps the five closest candidates and selects

$$i^\dagger = \arg\min_{i \in \mathrm{Top}_5^d(s)} |L(a_i) - L(h_s)|, \qquad L(a) = \sum_{t=1}^{T-1} \|a(t+1) - a(t)\|_2. \tag{5}$$

A separate soft-target variant activates the additional ListMLE loss (Xia et al., 2008) over these five candidates. Table 32 distinguishes target, support and objective changes. The safety term and fusion remain fixed in the primary comparison; Section 5.3 and Appendix O ablate them separately.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Data and evaluation. NAVSIM (Dauner et al., 2024; Cao et al., 2025) builds on logged nuPlan scenes (Caesar et al., 2021). We use 87,392 training, 5,278 validation and 12,146 navtest scenes. Validation shares training logs: 93.3% of its scenes lie within 1 s of a training scene. It selects checkpoints and the primary temperature. A separate log-disjoint retraining control tests this dependence (Table 3); official test logs are disjoint from both constructions. Primary rows share candidate pools and inputs (Section 4.2). Navtest uses official one-stage EPDMS; navhard uses official two-stage aggregation over 450 first-stage and 5,462 synthetic second-stage observations. Synthetic frames lack LiDAR files, so every row uses the same empty-LiDAR convention for the frozen planner. This convention limits comparison with published systems.

Training and uncertainty. Each main target uses three training seeds, AdamW (Loshchilov & Hutter, 2019), learning rate and weight decay 10^{-4}, batch size 48, and at most eight epochs. Checkpoints maximize cached validation PDMS, with patience three. Test scores are never used to select checkpoints or the primary temperature. Sample SD measures seed variation on fixed candidates. Navtest confidence intervals resample 136 logs of seed-averaged scene differences; they exclude full retraining and candidate-pool uncertainty (Appendix B). Differences are in pp.

Table 2: **RethinkDrive improves DDv2; target concentration changes the split ranking.** Official EPDMS $\times 100$, mean $\pm$ SD over three training seeds; Δ is relative to native selection before rounding. Bold highlights our fixed configuration, not column maxima or statistical significance. Learned rows share support 800, architecture and losses. Context is constant in training and recorded here; navhard shares the synthetic-frame sensor convention.

	navtest		navhard	
Selection / supervision	EPDMS ↑	Δ	EPDMS ↑	Δ
Planner, coarse score only	81.7	-2.0	21.1	$+0.3$
Planner, native fine selection	83.7	—	20.8	—
Metric-derived target (argmax)	74.5 ± 0.8	-9.2	10.7 ± 0.6	-10.1
Metric-derived target, tie-uniform	74.3 ± 1.2	-9.5	10.4 ± 0.7	-10.4
Metric target, reference tie-break	79.1 ± 1.6	-4.6	15.0 ± 0.8	-5.8
RethinkDrive (ours)	$\mathbf{87.6 \pm 0.1}$	$\mathbf{+3.9}$	$\mathbf{22.3 \pm 0.3}$	$\mathbf{+1.6}$
Soft reference ($\tau = 1$, validation choice)	83.0 ± 0.1	-0.7	22.3 ± 0.2	$+1.5$
Soft reference ($\tau = 0.3$)	84.2 ± 0.1	$+0.5$	24.3 ± 0.1	$+3.5$
Soft reference ($\tau = 0.1$)	87.0 ± 0.3	$+3.3$	23.4 ± 0.4	$+2.6$

5.2 TARGET EFFECTS AND NATIVE-BASELINE BENEFIT

Higher label scores do not predict a better learned selector. Metric-maximum labels have higher cached validation PDMS than reference labels (0.996 versus 0.915), yet the reference selector scores 13.1 pp higher on navtest. The label and deployment scores use different protocols; the contrast shows that ranking labels by their cached score does not predict the ranking of the trained selectors.

The benefit to the existing planner. RethinkDrive gains 3.93 ± 0.12 pp over native selection on navtest, with conditional 95% interval $[3.43, 4.44]$ pp. Removing EC and renormalizing the remaining weights changes the contrast to -0.49 pp, interval $[-1.02, 0.03]$; this is not evidence of equivalence. Against metric-target training, the EC-excluded gain is 7.53 pp, interval $[6.55, 8.58]$. Thus the native-relative gain is concentrated in temporal comfort, while the metric-target deficit extends beyond it. The navhard combined-score gain is 1.56 pp; concentrated soft targets score higher.

Training and input checks. No primary run diverges. The metric selector's unfused argmax falls in the validation tied set in 64% of scenes, versus 21% by chance; fitting is incomplete. Repeating all twelve runs with constant context throughout preserves the ordering and reference gains of 3.90 pp on navtest and 1.71 pp on navhard. This controls input mismatch without establishing optimal fitting.

5.3 MODEL SELECTION, FUSION AND TARGET CONTROLS

Table 3 adds controls planned after the primary results, with rules fixed before the new runs. They test validation dependence and alternatives to learned selection. They are reported separately from the primary protocol in Table 2.

Table 3: **The DDv2 gain survives new validation logs, penalty removal and rule comparisons.** EPDMS $\times 100$; learned rows report three-seed mean $\pm$ SD. Retraining uses a new log-disjoint split and constant context; $\lambda = 1$, $\beta = 0.3$ unless stated. Δ is relative to native. Validation selects $\tau = 1$; other temperatures are sensitivity results. Rule and fusion weights are selected on the new validation set. Bold highlights our fixed recipe, not column maxima.

DDv2 selection / target	navtest ↑	Δ (pp)	navhard ↑
Planner, native selection	83.71	—	20.79
Coarse selection / selected length rule	81.69	-2.02	21.09
Previous-plan rule ($\lambda = 0.3$)	81.35	-2.36	—
Metric maximum	74.17 ± 0.34	-9.55	10.96 ± 0.50
RethinkDrive (ours)	$\mathbf{87.64 \pm 0.07}$	$\mathbf{+3.93}$	$\mathbf{22.88 \pm 0.17}$
Soft reference ($\tau = 1$, selected)	82.84 ± 0.06	-0.87	22.40 ± 0.09
Soft reference ($\tau = 0.3$)	84.15 ± 0.19	$+0.44$	23.90 ± 0.24
Soft reference ($\tau = 0.1$)	87.23 ± 0.64	$+3.52$	22.82 ± 0.87
RethinkDrive, no penalty ($\beta = 0$)	87.57 ± 0.12	$+3.85$	22.59 ± 1.40
RethinkDrive, selected fusion ($\lambda = 0.3$)	87.93 ± 0.08	$+4.21$	—

Log-disjoint retraining preserves the target contrast. The new split assigns 85,077 scenes in 1,072 logs to training and 7,593 scenes in 119 logs to validation, with zero token or log overlap. All targets are retrained from random initialization with constant context throughout. Reference supervision retains $+3.93$ pp over native on navtest (95% interval $[3.39, 4.46]$) and gains 2.09 pp on navhard. The same cached-validation rule again selects $\tau = 1$, although concentrated soft targets score higher on test. This is a robustness check, not a causal estimate of validation leakage: the training composition also changes.

Training-free rules do not reproduce the learned gain. We penalize the coarse score by normalized trajectory length or by disagreement with the previously executed plan. The latter uses additional plan history and ego poses. Log-disjoint validation selects weights 0 and 0.3, respectively: length reduces to coarse selection, while consistency scores 81.35, below coarse (81.69) and native. Reference supervision scores 6.29 pp above this rule, interval $[5.61, 6.99]$. Validation-selected fusion ($\lambda = 0.3$) reaches 87.93; progress and compliance costs remain. Full grids and rule definitions are in Appendix O.

Tie resolution and metric alignment give partial recoveries. Within metric maxima, reference tie-breaking raises the primary score from 74.5 to 79.1, still below native and unrestricted reference supervision. A matched 20,279-scene pilot raises metric-target performance from 75.15 to 78.82 with aligned labels, below reference (87.33). The aligned score omits two-frame EC. Neither intervention identifies the remaining cause (Appendices L and M).

The target contrast survives removing the auxiliary penalty. Removing it changes log-disjoint navtest performance by -0.07 pp, interval $[-0.17, 0.03]$, with larger navhard seed variation. Thus this penalty is not required for the observed navtest gain. Table 32 also varies support, length matching, objective and temperature; these comparisons bound which components explain the benefit.

Target concentration trades performance across splits. Under log-disjoint retraining, soft $\tau = 0.1$ trails reference by 0.41 pp on navtest, interval $[-0.71, -0.10]$, while $\tau = 0.3$ leads it on navhard (23.90 versus 22.88). The hard target therefore retains no uniform advantage across splits; the preselected $\tau = 1$ remains the primary soft comparison.

5.4 VISUALIZING SELECTION AND BEHAVIOR

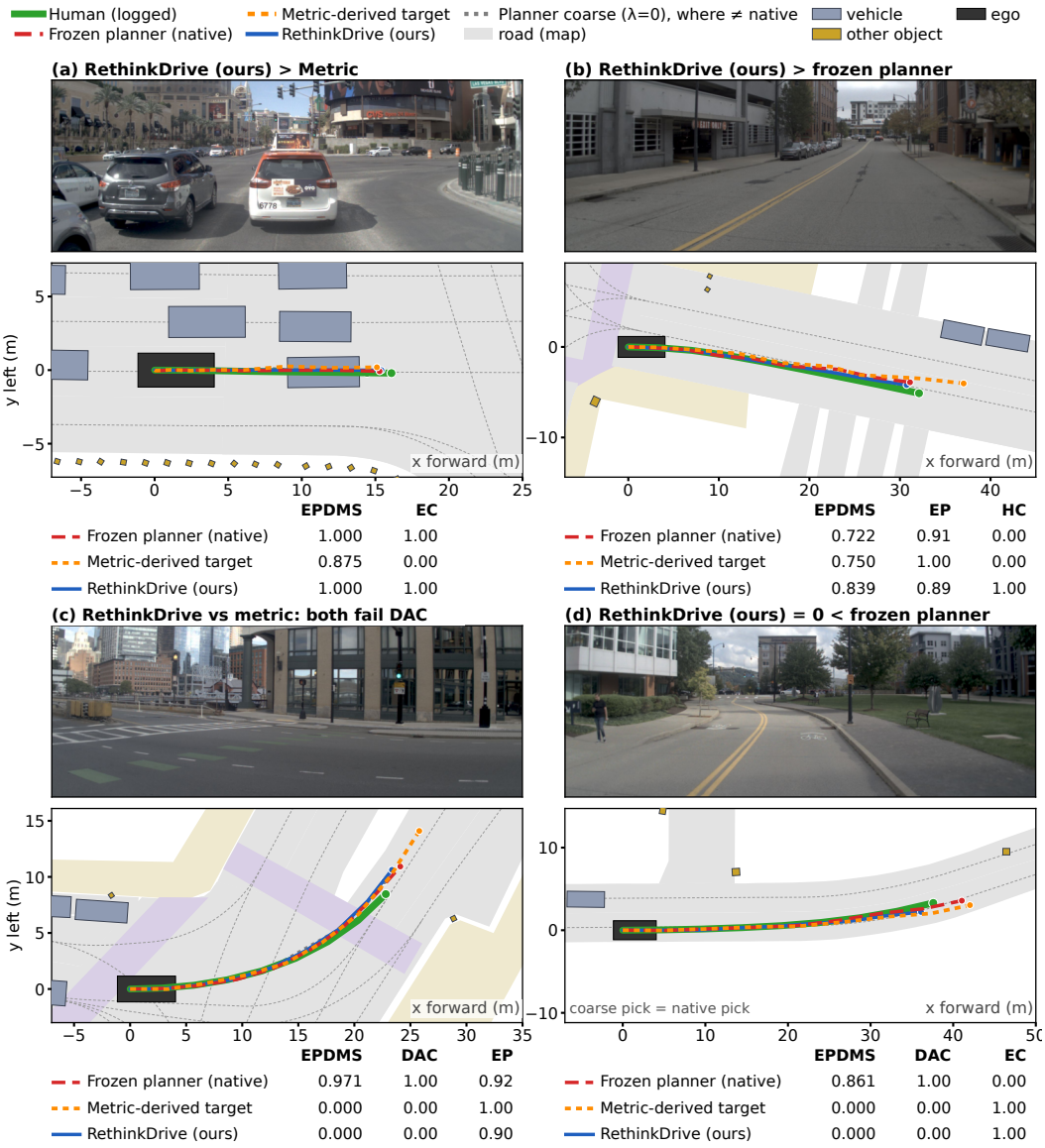

Figure 3: **Selection gains and failures under the primary protocol.** Four fixed-pool navtest scenes use official one-stage scores, seed 1 and recorded context. (a) Higher EC than metric-target selection. (b) Higher history comfort outweighs lower progress relative to native. (c–d) DAC failures despite different paths or high EC. Scenes follow a median-difference rule (Appendix D); proximity to a demonstration does not establish safety.

The learned ranking improves DDv2 before fusion. Without fusion, reference supervision scores 86.44 ± 0.06, 2.73 pp above native; fusion adds 1.21 pp. Metric-target ranking scores 62.13 ± 2.63. Soft $\tau = 1$ scores 87.18 ± 0.03 unfused and 83.0 fused, reversing its ordering with the hard target under the same checkpoints (Appendix O).

Scene-level outcomes. For seed 1, RethinkDrive improves 49.6% of scenes, ties on 23.3% and worsens 27.1% relative to native. Across seeds, 293–328 scenes newly score zero and 234–241 recover; new zeros exceed recoveries in every seed. DAC accounts for 207–234 new-zero cases (Appendix J). These are benchmark failures, not necessarily collisions.

5.5 SUPERVISION EFFECTS UNDER SCORE FUSION

Table 4 separates learned rankings from fused execution and its behavioral costs. Intervals condition on trained checkpoints and fixed candidate pools.

Table 4: **Fusion reverses target rankings or compresses their differences.** Navtest scores $\times 100$; learned rows average three seeds. *Selector* uses ranking alone; *Final* uses native selection or score fusion. Δ and log-bootstrap 95% intervals compare Final with native. EC, EP and DAC describe final execution; higher is better. Frozen coarse-only scores are $81.69/86.82$ for DDv2/DriveSuprim. EC uses $10{,}040$ adjacent-frame scenes; other scores use $12{,}146$ scenes. Bold highlights ours.

Selection / target	Selector	Final	Δ	95% interval	EC	EP	DAC
DiffusionDriveV2: fixed 800*-candidate pool*							
Planner, native selection	—	83.7	—	—	34.0	91.2	97.8
Metric maximum	62.1	74.5	-9.2	$[-9.91, -8.46]$	14.3	95.2	95.0
Tie-uniform	61.4	74.3	-9.5	$[-10.17, -8.77]$	13.2	95.5	95.1
RethinkDrive (ours)	**86.4**	**87.6**	$\mathbf{+3.9}$	$[\mathbf{3.43}, \mathbf{4.44}]$	**79.1**	**87.8**	**96.7**
Soft reference ($\tau = 1$)	87.2	83.0	-0.7	$[-1.09, -0.39]$	29.6	89.5	97.8
*DriveSuprim: fixed top-*256 *vocabulary shortlist*							
Planner, native selection	—	86.5	—	—	77.3	88.3	96.5
Metric maximum	80.4	86.7	$+0.2$	$[-0.33, 0.69]$	76.6	83.3	97.6
Tie-uniform	72.3	86.4	-0.2	$[-0.65, 0.35]$	75.4	84.0	97.2
RethinkDrive (ours)	**81.8**	**86.8**	$\mathbf{+0.3}$	$[\mathbf{-0.18}, \mathbf{0.81}]$	**79.5**	**84.0**	**97.1**
Soft reference ($\tau = 1$)	83.4	86.9	$+0.4$	$[-0.11, 0.91]$	78.4	83.8	97.4

The DDv2 gain is concentrated in temporal comfort. Within the official denominator, EC contributes $+4.36$ pp and the remaining term -0.42 pp; renormalizing without EC gives -0.49 pp. This is arithmetic, not causal attribution. Across $10{,}040$ frame pairs, RMS acceleration difference falls from 0.454 to $0.235\,\mathrm{m/s^2}$ without a temporal loss; metric and tie-uniform retain $0.729/0.674$. Progress and compliance decrease (Table 4); full sub-scores and temporal diagnostics are in Appendix K.

DriveSuprim exposes a different behavioral trade-off. Reference ranking alone raises EC from 77.3 to 81.0, but lowers DAC from 96.5 to 91.9 and the non-EC score from 88.1 to 82.9, yielding 81.83 overall. Fusion restores 86.83, near coarse-only 86.82, while changing 45.7% of coarse choices. Across four targets, fusion compresses the spread of means from 11.10 to 0.54 pp; all native-relative intervals include zero. The near-baseline totals therefore hide distinct learned behaviors: temporal comfort alone does not offset their other costs, and fusion restores performance. Candidate support and score scaling differ across planners (Yao et al., 2026); their causal roles remain unseparated (Appendices M and O).

Limitations. Gains hold on one of two planners. Our DDv2 checkpoint and low native EC are not matched to the published system (Appendix F). Fixed pools, partial metric alignment and limited training restrict generalization. Log-disjoint robustness does not identify a causal mechanism. Navhard uses empty LiDAR on synthetic frames; matched closed-loop CARLA/Bench2Drive controls remain untested (Dosovitskiy et al., 2017; Jia et al., 2024).

6 CONCLUSION

RethinkDrive evaluates supervision through both learning and execution. Reference supervision improves a frozen DDv2 planner, survives log-disjoint retraining and exceeds two training-free rules, primarily through temporal comfort with progress and compliance costs. Tie resolution only partly repairs metric supervision. Fusion reverses hard–soft rankings on DDv2 and compresses target differences on DriveSuprim. Reporting learned and fused selection separately exposes these effects when choosing a supervision target.

Ethics statement

We use public driving benchmarks and collect no new human-subject data. Simulation gains coexist with new failures. Demonstration distance is a supervision signal, not a safety measure. These results do not establish readiness for on-road deployment.

Reproducibility statement

Sections 4 and 5.1 specify targets, objectives and training. The supplement provides code, launch scripts, configurations, exact split lists, per-seed results, scene analyses and twelve selected selector checkpoints. Navtest is re-aggregated from scene outputs; navhard combined scores are checked against original scorer summaries, not independently re-aggregated. Retraining and simulation require external weights, data and caches. Appendix B discloses primary validation overlap; Appendices N and O document execution checks and log-disjoint retraining.

AI use statement

AI tools assisted with implementation, data processing, result aggregation and interpretation, experimental feedback, figures, and writing. The authors originated the research questions, directed experiments, and made final methodological and claim decisions. The authors are responsible for verifying all evidence and AI-assisted text, code, figures, references and results.

References

Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst. In *RSS*, 2019.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank using Gradient Descent. In *ICML*, 2005.

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

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

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

Sergio Casas, Abbas Sadat, and Raquel Urtasun. MP3: A Unified Model To Map, Perceive, Predict and Plan. In *CVPR*, 2021.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. MultiPath: Multiple Probabilistic Anchor Trajectory Hypotheses for Behavior Prediction. In *CoRL*, 2020.

Dian Chen and Philipp Krähenbühl. Learning From All Vehicles. In *CVPR*, 2022.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by Cheating. In *CoRL*, 2020.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end Driving via Conditional Imitation Learning. In *ICRA*, 2018.

Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with Misconceptions about Learning-based Vehicle Motion Planning. In *CoRL*, 2023.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-Driven Non-Reactive Autonomous Vehicle Simulation and Benchmarking. In *NeurIPS*, 2024.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *CoRL*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. ST-P3: End-to-end Vision-based Autonomous Driving via Spatial-Temporal Feature Learning. In *ECCV*, 2022.

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

Zhiyu Huang, Peter Karkus, Boris Ivanovic, Yuxiao Chen, Marco Pavone, and Chen Lv. DTPP: Differentiable Joint Conditional Prediction and Cost Evaluation for Tree Policy Planning in Autonomous Driving. In *ICRA*, 2024a.

Zhiyu Huang, Haochen Liu, Jingda Wu, and Chen Lv. Differentiable Integrated Motion Prediction and Planning With Learnable Cost Function for Autonomous Driving. *IEEE Transactions on Neural Networks and Learning Systems*, 2024b.

Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden Biases of End-to-End Driving Models. In *ICCV*, 2023.

Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think Twice before Driving: Towards Scalable Decoders for End-to-End Autonomous Driving. In *CVPR*, 2023.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-To-End Autonomous Driving. In *NeurIPS*, 2024.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. In *ICLR*, 2025.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In *ICCV*, 2023.

Bo Jiang, Shaoyu Chen, Hao Gao, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. VADv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning. In *ICLR*, 2026.

Jungho Kim, Jiyong Oh, Seunghoon Yu, Hongjae Shin, Donghyuk Kwak, and Jun Won Choi. SafeDrive: Fine-Grained Safety Reasoning for End-to-End Driving in a Sparse World. In *CVPR*, 2026. arXiv:2602.18887.

Kailin Li, Zhenxin Li, Shiyi Lan, Yuan Xie, Zhizhong Zhang, Jiayi Liu, Zuxuan Wu, Zhiding Yu, and Jose M. Alvarez. Hydra-MDP++: Advancing end-to-end driving via expert-guided hydra-distillation. *arXiv preprint arXiv:2503.12820*, 2025a.

Yaguang Li, Jiaru Zhang, Chuheng Wei, Can Cui, and Ziran Wang. Designing Versatile Samples for Learned Trajectory Scoring. *arXiv preprint arXiv:2609.01799*, 2026.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing End-to-End Autonomous Driving with Latent World Model. In *ICLR*, 2025b.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, Yu-Gang Jiang, and Jose M. Alvarez. Hydra-MDP: End-to-end Multimodal Planning with Multi-target Hydra-Distillation. *arXiv preprint arXiv:2406.06978*, 2024.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In *ECCV*, 2022.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving. In *CVPR*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

Tung Phan-Minh, Elena Corina Grigore, Freddy A. Boulton, Oscar Beijbom, and Eric M. Wolff. CoverNet: Multimodal Behavior Prediction Using Trajectory Sets. In *CVPR*, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*, 2023.

Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A. Sophia Koepke, Zeynep Akata, and Andreas Geiger. PlanT: Explainable Planning Transformers via Object-Level Representations. In *CoRL*, 2023.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *AISTATS*, 2011.

Shuyao Shang, Yuntao Chen, Yuqi Wang, Yingyan Li, and Zhaoxiang Zhang. DriveDPO: Policy Learning via Safety DPO For End-to-End Autonomous Driving. In *NeurIPS*, 2025.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. SparseDrive: End-to-End Autonomous Driving via Sparse Scene Representation. In *ICRA*, 2025.

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. SparseDriveV2: Scoring is All You Need for End-to-End Autonomous Driving. *arXiv preprint arXiv:2603.29163*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NeurIPS*, 2017.

Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided Control Prediction for End-to-end Autonomous Driving: A Simple yet Strong Baseline. In *NeurIPS*, 2022.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise Approach to Learning to Rank: Theory and Algorithm. In *ICML*, 2008.

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. GoalFlow: Goal-Driven Flow Matching for Multimodal Trajectories Generation in End-to-End Autonomous Driving. In *CVPR*, 2025.

Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M. Alvarez, and Zuxuan Wu. DriveSuprim: Towards Precise Trajectory Selection for End-to-End Planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2506.06659.

Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-To-End Interpretable Neural Motion Planner. In *CVPR*, 2019.

Yinan Zheng, Ruiming Liang, Kexin Zheng, Jinliang Zheng, Liyuan Mao, Jianxiong Li, Weihao Gu, Rui Ai, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Diffusion-Based Planning for Autonomous Driving with Flexible Guidance. In *ICLR*, 2025.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. DiffusionDriveV2: Reinforcement Learning-Constrained Truncated Diffusion Modeling in End-to-End Autonomous Driving. *arXiv preprint arXiv:2512.07745*, 2025.

A SCORES AND NOTATION

The target-construction score, the navtest evaluation score and the navhard evaluation score serve different purposes. Keeping them separate is essential to interpreting the comparisons. NAVSIM supplies the simulation metrics (Dauner et al., 2024); its pseudo-simulation extension adds synthetic observations and a two-stage evaluation (Cao et al., 2025). Unless specified otherwise, bold method names and values highlight the fixed RethinkDrive configuration, not column maxima or statistical significance.

Table 5: Notation used in the main text.

Symbol	Meaning
s, a_i, K	Scene, candidate trajectory, and number of candidates (800 for DDv2; 256 for the DriveSuprim shortlist).
h_s, d_i	Recorded human trajectory and candidate's mean displacement from it.
f_{base}, z_θ	Frozen coarse score and added selector's raw logit.
π_θ, q	Training prediction $\mathrm{softmax}(z_\theta)$ and target distribution.
σ, λ	Logistic sigmoid and inference fusion weight (1).
m, E	Cached PDMS and current extended evaluation score.
NC, DAC	No at-fault collision and drivable-area compliance.
DDC, TLC	Driving-direction compliance and traffic-light compliance.
EP, TTC	Ego progress and time-to-collision within bound.
C, HC, EC	Comfort, history comfort, and extended comfort.
LK	Lane keeping.

Cached score for training. Target builders use cached per-candidate sub-scores, composed as

$$m = \mathrm{NC}\,\mathrm{DAC}\,\frac{5\mathrm{TTC} + 5\mathrm{EP} + 2\mathrm{C}}{12}. \tag{6}$$

This is the cached PDMS used for the metric-maximum and uniform-maxima targets and for checkpoint selection. It is not the extended evaluation score. The sub-scores used by the common safety penalty are also cached simulation labels, whereas the score vector supplied to the selector as an input is predicted by the frozen planner.

Current navtest evaluation. Let $G = \mathrm{NC}\,\mathrm{DAC}\,\mathrm{DDC}\,\mathrm{TLC}$. The upstream one-stage evaluator uses

$$E = G\,\frac{5\mathrm{EP} + 5\mathrm{TTC} + 2\mathrm{LK} + 2\mathrm{HC} + 2w_{\text{adj}}\mathrm{EC}}{14 + 2w_{\text{adj}}}, \tag{7}$$

where $w_{\text{adj}} = 1$ when a preceding frame in the same log is available within 0.55 s, supporting the two-frame extended comfort calculation, and 0 otherwise. Exactly $10{,}040$ navtest scenes have EC and $2{,}106$ do not, with the same partition across methods. The denominator is 16 with EC and 14 without it; missing EC is not scored as zero. The replay agent returns stored predictions without loading sensors. All $12{,}146$ filtered navtest observations are scored. The cache also contains 65 observations outside this filter; these are excluded.

Training metric and pose convention. Cached training PDMS uses the formula above, but pads eight predicted poses to ten by repeating the final pose. Test trajectories contain eight poses at 0.5 s intervals. The training labels and test score therefore differ in both metric definition and pose convention. The matched-subset pilot in Appendix M intervenes on label construction by removing padding and using the single-stage evaluation criteria without EC. It retains cached-PDMS checkpoint selection, so it is a partial alignment control rather than a fully metric-aligned recipe.

navhard protocol. The two-stage evaluator aggregates 450 first-stage and $5{,}462$ synthetic second-stage observations using weights determined by first-stage predictions. Runs invoking the local single-stage fallback are invalid for this comparison. All synthetic second-stage observations lack LiDAR files in the supplied release. The local loader supplies empty LiDAR to the camera+LiDAR planner for these observations, identically across baselines and selectors. This is a local input convention, not a verified reproduction of the published system. Table 2 reports the matched

Table 6: **Targets.** Target properties on hard-val (n=5,278) with one displacement definition; "cached score" is the cached PDMS of the target candidate, not the evaluation score. All configurations score all 800 candidates at inference.

target	distribution q	support	c_{rank}	mean d (m)	cached score
Metric maximum	one-hot, first tied PDMS maximum	800	0	2.545	0.996
Uniform maxima	uniform over tied PDMS maxima	800	0	—	0.996
RethinkDrive (ours)	one-hot, $\arg\min \mathbf{d_i}$	**800**	**0**	**0.116**	**0.915**
Soft reference	$\mathrm{softmax}(-d_i/\tau)$, τ=1	800	0	—	—
ablation variants					
Nearest (80)	one-hot, $\arg\min d_i$	first 80	0	0.162	0.908
Nearest (80) + length	one-hot, Eq. 5	first 80	0	0.188	0.904
Soft + ranking	$\mathrm{softmax}(-d_i/\tau)$, τ=1	800	**1**	—	—

four-target navhard comparison under this shared convention. It does not establish equality with sensor-complete evaluation.

B TRAINING AND EVALUATION DETAILS

Data and candidate support. Training uses 87,392 cached scenes; hard-val contains 5,278 scenes held out by token, not by log. All training targets refer to the same cached 800 candidates and predicted features. The first 200 candidates are 20 anchors under 10 diffusion-noise draws. Three groups of randomly rescaled copies provide the remaining 600. Restricting a target to the first 80 therefore selects a structured subset of the unperturbed samples, not a random subset of the full pool. Every trained selector still scores all 800 candidates at inference.

Validation dependence and selection. Training and validation contain no identical scene tokens, but all 979 validation logs also contain training scenes. The split audit finds a training scene from the same log within 1 s for 93.3% of validation scenes; the median separation is 0.50 s. Validation therefore does not estimate generalization to unseen logs and may give an optimistic selection signal. It is used for cached-PDMS checkpoint selection, early stopping and primary soft-temperature selection. Validation candidate statistics elsewhere in the paper are descriptive pool diagnostics. The official navtest and navhard test logs are disjoint from training. Their reported scores remain measurements of the selected checkpoints. The separate log-disjoint retraining study in Table 3 repeats checkpoint and temperature selection, retaining the reference–metric contrast and the selected $\tau = 1$. Because training membership also changes, it measures sensitivity rather than a causal leakage effect. The supplement releases exact scene lists, overlap summaries and the new split rule.

Model and optimisation. The selector uses two attention layers with hidden width 128. Candidate inputs consist of 16 geometry features and 6 predicted base scores; scene slots hold the driving command and four ego-state values, but they are constant in training as detailed below. The reserved visual slot is zero-filled. There are 0.5193 million parameters in the measured module. Training uses AdamW (Loshchilov & Hutter, 2019), learning rate 10^{-4}, weight decay 10^{-4}, batch size 48, and a maximum of eight epochs. Early stopping uses patience three on the cached score of the selected candidate on hard-val. Training uses softmax, while inference uses sigmoid fusion (Eqs. 2 and 3). Appendix O tests removing fusion and selecting its weight; the nonlinear link itself is not compared with alternatives.

Additional ranking term. Only Soft + ranking activates the extra ListMLE term (Xia et al., 2008). Let $j_1, \ldots, j_5$ be the five smallest-displacement candidates among the first 80, ordered by displacement. The implemented term is

$$\mathcal{L}_{\text{rank}} = \sum_{r=1}^{5} \left[\log \sum_{u=r}^{5} \exp z_\theta(a_{j_u}, s) - z_\theta(a_{j_r}, s) \right]. \tag{8}$$

All four main targets set its coefficient to zero. The shared safety penalty is an expectation under the differentiable softmax distribution; it does not backpropagate through a discrete argmax. No

language-model teacher or policy-gradient objective is active in the four primary target configurations. The broader comparison separately includes a policy-gradient control driven by metric reward.

Fixed candidates and baselines. A cache stores one 800-candidate pool and the frozen predicted features per observation. Every learned target reuses that cache. A deterministic seed derived from the scene identifier controls generation; the accompanying manifest links the shared cache to the run summaries. The native baseline is the frozen checkpoint's own coarse-to-top-32-fine selection path; the coarse baseline omits the fine scorer. The local checkpoint's equivalence to an official released checkpoint is not established. The recorded native prediction matches that checkpoint's selected output. The completed navtest candidate lookup finds a stored candidate for every native prediction, with no unmatched scenes in the $12{,}146$-scene audit.

Input conditions and scope. Training uses a constant straight-command vector $[0, 1, 0, 0]$ and zero ego state in all $87{,}392$ scenes, because the training loader's per-scene signal files were absent. All $5{,}278$ validation scenes have those files. Checkpoint selection and primary evaluation therefore use recorded commands and $[v_x, v_y, 0, 0]$. The earlier observed-acceleration condition instead uses $[v_x, v_y, a_x, a_y]$ at test time and remains secondary evidence below. Neither evaluation convention matches all training inputs. Geometry, predicted-score order, the reserved zero visual slot and the fusion rule are shared. Two controls distinguish inference sensitivity from a fully matched protocol. First, the twelve retained checkpoints are evaluated with constant context. Second, training and validation selection are repeated with constant context throughout. Both preserve the navtest target ordering; the latter shifts target means by at most 0.27 pp (Appendix G). The original primary rows are retained rather than replaced after observing these checks.

Training-state audit. Logs from the twelve primary runs show no NaN, infinity or numerical divergence. Stored-checkpoint forward passes put the metric selector's own argmax in the cached tied set on about 64% of validation scenes, compared with 67% for tie-uniform and 21% under uniform random choice. Agreement with the first maximizing index is only 5.9%; RethinkDrive matches its reference index on 4.9%. These measure different targets and do not rank learning quality. Set-level learning exceeds chance, while underfitting and gains from objective-specific tuning remain possible. The comparison uses a common budget rather than a best-effort search for every objective.

Broader configurations. The DPO-style row uses the first-80 plus length target against the remaining candidates and omits both the safety penalty and ListMLE. The policy-gradient row also omits them. The hybrid adds sub-score regression to that first-80 plus length target, retaining the safety penalty. Sub-score-only distillation fuses the coarse score with a composition of predicted EPDMS components, rather than the scalar selector logit. These four rows each use one retained checkpoint, trained with three-device data parallelism at total batch 48; two have no recorded seed. Main target rows use three one-device runs at batch 48. Thus these broader rows change more than one component and are not matched objective ablations.

Conditional confidence intervals. The confidence-interval analysis resamples 136 scene logs with replacement, keeping scenes within each sampled log together. For each of $10{,}000$ draws it averages seed-averaged per-scene score differences; percentile endpoints give the reported 95% intervals. These intervals describe log resampling conditional on the retained models and candidate pool, not full training-and-pool uncertainty. The aggregate means and sample SD can be recomputed from the accompanying per-seed summaries; reproducing the intervals also requires the referenced per-scene outputs.

Temperature selection. Seed-1 hard-val scores for $\tau = 0.1, 0.3, 1.0$ m are 0.938468, 0.944553 and 0.947301, respectively. The largest validation score selects 1.0 m. The primary comparison retains this choice. All three temperatures have three training seeds. On navtest, 0.3 and 0.1 reach 84.17 ± 0.09 and 86.99 ± 0.33; on navhard they reach 24.31 ± 0.14 and 23.42 ± 0.41. Cached validation ranking does not transfer to either test split. Appendix L reports the hard–soft contrasts.

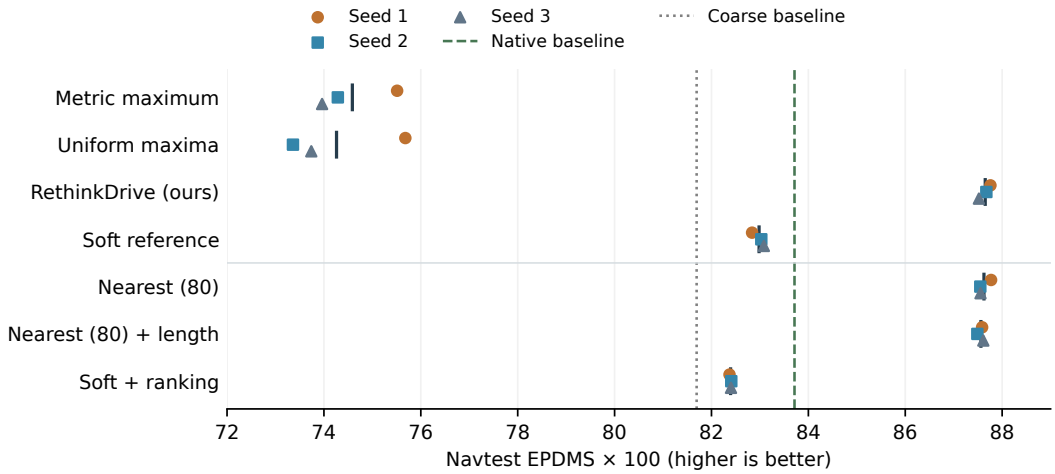

Figure 4: **Secondary input-condition results for individual training seeds.** Observed-acceleration navtest, upstream one-stage EPDMS. Markers denote seeds and a short vertical line denotes their mean. Separate reference lines show the coarse and native baselines.

Table 7: **Secondary observed-acceleration navtest scores.** Observed-acceleration condition, upstream one-stage EPDMS on the 0–1 scale. Native baseline: 0.837137; coarse: 0.816937. SD uses three independent training seeds, conditional on the fixed candidate cache.

Target	Seed 1	Seed 2	Seed 3	Mean	SD
Metric maximum	0.755109	0.742870	0.739621	0.745867	0.008167
Uniform maxima	0.756805	0.733607	0.737392	0.742601	0.012445
RethinkDrive (ours)	**0.877583**	**0.876740**	**0.875205**	**0.876509**	**0.001206**
Soft reference	0.828382	0.830283	0.830811	0.829825	0.001278
Nearest (80)	0.877716	0.875489	0.875537	0.876247	0.001272
Nearest (80) + length	0.875864	0.874881	0.876116	0.875620	0.000653
Soft + ranking	0.823722	0.824086	0.824029	0.823946	0.000196

C Details of the tie diagnosis

For a scene with maximum score $m_{\max}$, the audit defines $\mathcal{T}(s) = \{i : m_i \geq m_{\max} - 10^{-6}\}$. Saturation means $m_{\max} \geq 1 - 10^{-6}$; a tie means $|\mathcal{T}(s)| \geq 2$. The conditional median in Table 1 uses only tied scenes. Table 8 also gives the overall median, which includes unique maxima as sets of size one, and both interquartile ranges. The first exact maximum supplies the metric-maximum label; Uniform maxima assigns mass $1/|\mathcal{T}(s)|$ to each member of the audit set. These are different constructions when distinct scores are within the tolerance.

Evaluation-aligned scoring. All 800 candidates in each of the 5,278 hard-val scenes were re-scored using the single-stage evaluation definition with eight poses and no padding. The two-frame extended comfort term is excluded. Saturation remains 82.5%, and 84.7% of scenes have tied maxima, with median tied-set size 98. The aligned and cached maximizing indices agree in 49.8% of scenes; the aligned maximum belongs to the old tied set in 70.7%. Mean displacement from the human reference is 2.106 m for the aligned argmax and 2.545 m for the cached argmax. The RethinkDrive target lies in the aligned tied set in 34.4% of scenes. Thus the alternative target is not generally a tie-break within the metric optimum. These measurements diagnose ambiguity. The separate matched-subset training comparison in Appendix M measures learned performance from aligned labels, without attributing the remaining deficit to ties.

Reference-target uniqueness. Using tolerance 10^{-6} m, minimum-displacement ties occur in one of 5,278 validation scenes and two of 4,370 sampled training scenes. A unique minimum is unchanged under candidate permutation. The current evidence does not establish strict order invariance for the

Table 8: **Saturation and ties in two candidate pools**, one script, one definition (tolerance 10^{-6}). Tied-set size counts candidates attaining the scene maximum; the *conditional* median is over scenes with at least two tied candidates. Split, metric and storage precision differ between pools, so rows are not ranked against each other.

pool	K	split, scenes	score	max = 1	tie	tied-set size, median [IQR]	
						overall	conditional
DiffusionDrive-v2 candidates	800	hard-val, 5,278	PDMS	88.2%	85.3%	87 [9–310]	134 [27–360]
DiffusionDrive-v2 candidates	800	train, 87,392	PDMS	88.1%	85.3%	93 [10–317]	145 [29–361]
Same candidates, aligned score	800	hard-val, 5,278	EPDMS (no EC)	82.5%	84.7%	—	98
DriveSuprim vocabulary	8192	navtrain, 103,288	EPDMS	91.8%	90.5%	183 [32–475]	223 [68–523]

rare ties, so the method is described as nearly unique rather than as an unconditional order-invariant selection rule.

Storage precision. DiffusionDrive-v2's PDMS is composed from cached single-precision sub-scores. DriveSuprim's public cache stores half-precision sub-scores; we compose its single-stage score in double precision as $G(5\mathrm{EP} + 5\mathrm{TTC} + 2\mathrm{LK} + 2\mathrm{HC})/14$. This audit score is separate from the current navtest evaluation with temporal EC. Casting to double precision does not undo quantisation in the source values. Differences in data, score definition, precision and candidate generation preclude ranking the two planners by their tie rates.

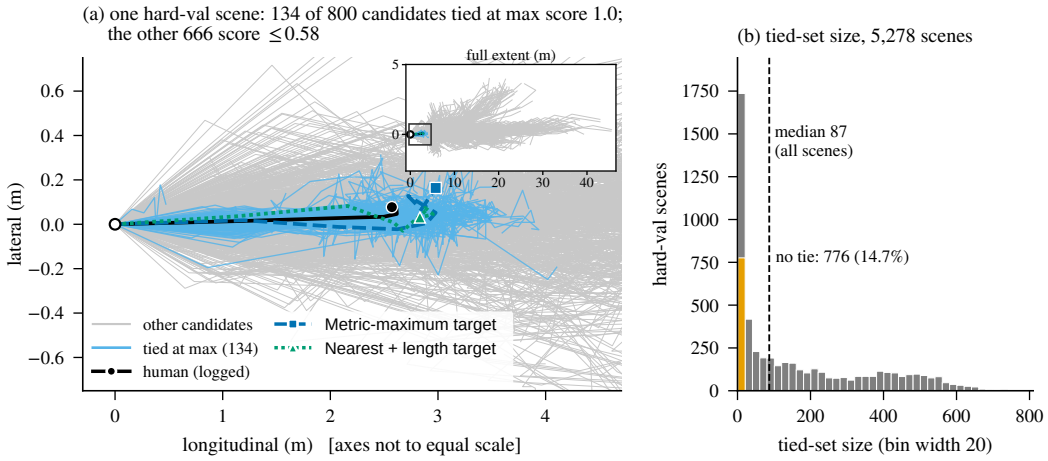

Figure 5: **What a tie means.** (a) The tied hard-val scene whose tied-set size is closest to the median over tied scenes (134; lexicographic tie-break). 134 of its 800 candidates score the cached-PDMS maximum 1.0 and the rest at most 0.58. The metric target is the first tied index; the Nearest (80) + length target (Section 5.3) lies inside the tied set here. The main view is zoomed and not to equal scale; the inset shows the full extent. (b) Tied-set size over all 5,278 hard-val scenes; 776 (14.7%) have a unique maximum.

Trajectory illustration in the main text. Figure 1 plots one validation scene from a retained candidate-pool visualization. Both panels contain the same 800 candidates; 372 attain the cached maximum. The selected metric label scores 1.0 and lies 1.13 m from the demonstration, while the nearest label scores 0.0 and lies 0.11 m away. These distances measure agreement, not safety. The lower strip reports aggregate navtest results and is separate from these scene-level label properties.

Geometry of ties. In 60 tied sets drawn from the fixed 2,500-scene audit sample, no candidate pair is exactly duplicated and 0.78% of pairs are within 0.1 m. The mean endpoint spread is 1.251 m. These measurements establish geometric variation inside the audited tied sets, not that one tied candidate is safer than another.

Cached validation and test evaluation answer different questions. A separate diagnostic compares choices on identical cached hard-val candidates. Among 5,210 scenes where the Metric and RethinkDrive (ours) selectors disagree, RethinkDrive (ours) has a lower cached PDMS in 3,015, a higher one in 346 and the same score in 1,849. The test-set advantage under extended evaluation scores is therefore not an improvement in every metric. This also makes the cached-PDMS versus evaluation-score mismatch a substantive limit of the study rather than a naming detail.

D QUALITATIVE EXAMPLE SELECTION

Figure 3 uses recorded front-camera images, maps and executed trajectories from the same fixed-pool primary evaluation as Table 2. Learned selectors use seed 1 and recorded context. The native planner is the baseline for panels (b,d); the metric-derived selector is the comparison for (a,c). The coarse choice is shown only where it differs from native and is not the comparison baseline.

For each condition, scenes are ordered by distance from the median signed score difference. Ties use a stable scene-identifier ordering, skipping scenes already used. This is a deterministic rule, not selection by visual appearance; it does not guarantee representative road geometry. Identifiers, checkpoints and prediction provenance accompany the figure as data rather than labels on the plot.

Table 9: **Primary-protocol qualitative examples.** Official navtest EPDMS on the 0–1 scale. Condition sets overlap. Panels (c,d) fail DAC for both learned selectors.

Panel / selection condition	Scenes	Native	Metric	**RethinkDrive**
(a) Above metric-target selection	8040	1.0000	0.8750	**1.0000**
(b) Above native selection	6027	0.7220	0.7500	**0.8393**
(c) Different paths, similar scores	1113	0.9712	0.0000	**0.0000**
(d) Zero while native is positive	293	0.8609	0.0000	**0.0000**

Panel (c) requires the two learned selections' endpoints to differ by more than 2 m and their scores by less than 0.01. Both receive zero here, so the equal score reflects a gate failure, not ceiling saturation. In (d), EC passes but DAC fails. In (b), history comfort improves while progress decreases. None equates demonstration agreement with safety.

Table 10: **Secondary behavioral audit under observed acceleration.** Changes from native selection on navtest, averaged over three seeds. These are not the recorded-context primary results. "New zero" changes a positive score to zero; "recovered" is the reverse. EC is extended comfort on the 0–1 scale. Counts are rounded. Bold highlights the RethinkDrive row.

Target	Improved (%) ↑	Worse (%) ↓	New zero ↓	Recovered ↑	Δ EC
Metric maximum	27.1	44.3	900	116	−0.196
Uniform maxima	28.0	41.8	954	107	−0.208
RethinkDrive (ours)	**50.0**	**26.7**	**310**	**237**	**+0.450**
Soft reference	20.6	42.6	145	185	−0.044

E SCOPE OF EARLIER EXPERIMENTS

Earlier development experiments changed target support, auxiliary losses, backbones or evaluation paths together. They do not form additional replications of the matched comparisons in the main text. We retain the relevant negative findings and corrections here to delimit what the present study establishes.

Earlier NAVSIM configurations. Earlier single-run Nearest (80) and Nearest (80) + length results included navhard gains. Those configurations used an extra ranking term and an earlier evaluation pipeline. A subsequent comparison under the earlier protocol did not reproduce an improvement over the navhard baseline; the updated fixed-pool study reports navtest and navhard under the primary

Table 11: **Secondary observed-acceleration sub-score changes against native selection.** Observed-acceleration condition, mean over three training seeds, on the 0–1 scale. These are the measurements behind Table 10, separate from the recorded-context primary comparison. Positive values improve the named sub-score; multiplicative gating means these changes are not additive contributions to overall EPDMS.

Target	NC	DAC	EP	TTC	LK	HC	EC
Metric maximum	−0.027	−0.029	+0.040	−0.034	−0.015	−0.210	−0.196
Uniform maxima	−0.031	−0.028	+0.044	−0.040	−0.023	−0.162	−0.208
RethinkDrive (ours)	**−0.001**	**−0.011**	**−0.034**	**−0.001**	**+0.018**	**+0.063**	**+0.450**
Soft reference	+0.002	+0.000	−0.017	+0.002	−0.001	−0.002	−0.044

protocol. Its values replace those earlier comparisons; they are not pooled across evaluation paths. A previous comparison of dynamic time warping and mean displacement also used different candidate supports; it does not isolate the distance metric.

Cross-domain and closed-loop scope. In an exploratory Bench2Drive (Jia et al., 2024) test, NAVSIM-trained variants achieved at most one route with DS 100 out of ten. This corrects the earlier all-zero statement; completed evaluator status and DS 100 are different outcomes (Table 17). Separate TCP-based (Wu et al., 2022) experiments used a pool constructed by perturbing its single output, which changes the candidate-generation problem as well as the training domain. Their offline evaluation contained a horizon mismatch, and the closed-loop comparison lacked a reproduced TCP control on the same route set. We therefore do not use those experiments as evidence for cross-backbone improvement. Section H.1 reports the retained closed-loop aggregate, with no claim of baseline improvement or on-road deployment readiness.

Other explored directions. Teacher-based supervision, geometric gates and pool augmentation were evaluated in earlier configurations. They are not components of RethinkDrive as evaluated here. Likewise, a test-split weight sweep on another planner is not held-out evidence of generalisation. Including these exploratory results as additional confirmations would conflate different interventions. The main evidence is the two-pool diagnostic, the matched four-target comparison and the separately identified extended configurations, including their negative outcomes.

F Positioning against related planners

Table 13 summarizes the scope of the closest comparisons. The supervised signals overlap with established work. The experimental question here is their conversion into target distributions while holding a separate selector and its training recipe fixed.

Representation and waypoint refinement. MP3 combines an online map and predictions of agent states with neural motion planning (Casas et al., 2021). TransFuser fuses camera and LiDAR (Chitta et al., 2023); BEVFormer builds bird's-eye-view features (Li et al., 2022). ST-P3, UniAD and VAD connect perception and prediction to planning (Hu et al., 2022; 2023; Jiang et al., 2023). LAW learns future latent scene features (Li et al., 2025b), while SafeDrive models trajectory-conditioned interactions (Kim et al., 2026). These approaches change the information or representation available for planning. ThinkTwice uses predicted future context and retrieved features to refine coarse positions and actions through learned offsets (Jia et al., 2023). Here, refinement changes only the selected index among existing trajectories: neither the scene encoder nor the waypoints are updated.

Vocabulary scoring and target construction. Discrete trajectory supervision also predates end-to-end ego planning. CoverNet formulates multimodal behavior prediction as classification over a set of trajectories (Phan-Minh et al., 2020). MultiPath predicts probabilities over trajectory anchors and regresses waypoint offsets (Chai et al., 2020). Their task is forecasting behavior, whereas our study selects an ego trajectory to execute. These precedents mean that neither a discrete trajectory label nor a demonstration-based target is novel by itself. VADv2 learns a demonstration-supervised probability distribution over a discretized planning vocabulary (Jiang et al., 2026). SparseDrive combines a parallel motion planner with hierarchical selection and collision-aware rescoring (Sun et al., 2025).

SparseDriveV2 expands vocabulary coverage through factorized paths and velocity profiles and uses coarse-to-fine scoring (Sun et al., 2026). Our frozen-pool experiment does not test those architectures or establish superiority to them. It changes the supervision distribution within a shared selection pipeline, separating that change from vocabulary coverage and scorer capacity.

Scorer training on frozen planners. Li et al. (2026) train separate scorers on frozen DiffusionDrive and MeanFuser planners and augment their training pools with perturbations of recorded trajectories. Their intervention changes the training samples for component-score supervision. Our matched comparison retains the candidates and changes the target distribution. Frozen planner reuse and a lightweight scorer alone do not distinguish the present work.

Learned planning costs. NMP learns a spatial cost volume and selects the lowest-cost trajectory among physically feasible samples (Zeng et al., 2019). DIPP learns prediction and cost-function weights through a differentiable nonlinear trajectory optimizer (Huang et al., 2024). Both establish that demonstrations can supervise planning costs. Our experiment instead holds candidate geometry and scorer architecture fixed to examine how the target distribution affects a learned selector.

Ranking and soft targets. Pairwise probabilistic ranking is established by RankNet (Burges et al., 2005); ListMLE provides a listwise likelihood objective (Xia et al., 2008). Temperature-softened teacher distributions are also established in knowledge distillation (Hinton et al., 2015). Our soft-reference control constructs probabilities from demonstration distance rather than teacher logits. These connections position target construction within existing learning methods; neither a new ranking loss nor the general use of soft supervision is a contribution of this work.

Published DDv2 context. Table 12 separates the published NAVSIM v2 navtest result from our local experiment. The public checkpoint's identity and exact evaluation configuration have not been matched to the local run. Only the local native row supplies the baseline for the 3.9 pp improvement; the published row is contextual evidence, not a SOTA comparison. The same published table reports EC 91.0, whereas the local native run has EC 34.0. The cause of this difference has not been isolated across checkpoint, input and evaluation conditions. Since the local gain is concentrated in EC, it does not establish an analogous gain on the released system. DDv2 trains component scorers with binary cross-entropy and a continuous-score ranking loss; our metric-maximum classification control does not reproduce that native training objective.

Table 12: **Published and local navtest context.** All scores are EPDMS $\times 100$. The first four rows are transcribed from Table 2 of Zou et al. (2025), all ResNet-34; they are external results, not local reproductions.

System and evidence	EPDMS ↑	Comparison status
TransFuser (Chitta et al., 2023), reported by DDv2	76.7	External context only
Hydra-MDP++ (Li et al., 2025a), reported by DDv2	81.4	External context only
DriveSuprim (Yao et al., 2026), reported by DDv2	83.1	External context only
DiffusionDriveV2, published ResNet-34	85.5	External context only
Local frozen planner, native selection	83.7	Matched local baseline
RethinkDrive on the local frozen planner	87.6 ± 0.1	Three training seeds

Evaluation context. Logged datasets such as nuScenes (Caesar et al., 2020), closed-loop planning in nuPlan (Caesar et al., 2021), CARLA and Bench2Drive (Dosovitskiy et al., 2017; Jia et al., 2024), and NAVSIM's recorded-scene and pseudo-simulation protocols (Dauner et al., 2024; Cao et al., 2025) evaluate different properties. Published scores under these protocols do not supply a matched baseline for our local refinement runs.

G Exact selector features

The implemented feature function computes the following vector in ego coordinates. Let x_t, y_t, ψ_t denote candidate position and heading, and let $u_t = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2}$ be the step length. The 16 entries are:

Table 13: **Training and supervision in related planners**, following the original papers.

	what is trained	candidates	selection supervision
Hydra-MDP	backbone and scoring heads, end to end	fixed vocabulary	distance-based imitation target and per-metric score distillation
DriveSuprim	full planner with coarse-to-fine scorer	large vocabulary, progressively filtered	metric-derived scores, with augmentation and self-distillation
DriveDPO	planning policy	policy outputs	unified imitation/safety distribution, then safety preference optimisation
This work	a separate 0.52 M selector; planner frozen	800 samples per scene from a frozen generator	matched comparison of metric-derived and human-reference targets

1. x_T, y_T, $|y_T|$, $\max_t y_t$, $\max_t |y_t|$, $\mathrm{mean}_t|y_t|$;
2. $\mathrm{mean}_t u_t$, $\max_t |\psi_{t+1} - \psi_t|$, $\mathrm{mean}_t|\psi_{t+1} - \psi_t|$, $\mathrm{mean}_t|u_{t+1} - u_t|$;
3. $\max_{t>T/2} |y_t| - \max_{t\le T/2} |y_t|$, $\max_t |\psi_t|$, ψ_T, $\mathrm{sign}(y_T)\mathrm{sign}(\psi_T)$;
4. $\sum_t \mathbf{1}[u_t < 0.05]$, $\sum_t u_t$.

These are geometric descriptors, not physically calibrated curvature or jerk measurements. The command slot is four-dimensional. During training it is fixed to $[0, 1, 0, 0]$, with all four ego-state entries zero. Validation and primary evaluation supply recorded command and $(v_x, v_y, 0, 0)$; the secondary condition additionally supplies acceleration. Changing recorded context to training constants changes 6–12% of fused choices for the one-hot targets and about 1% for soft $\tau = 1$. Table 14 distinguishes test-only input substitution from repeated training and validation selection. The fully matched control uses constant command and ego state in training, validation and test, while keeping the architecture, candidate cache, targets and optimization budget fixed. All twelve selected epochs match their primary counterparts: metric $1/6/6$, tie-uniform $4/4/4$, reference $5/5/5$, and soft $7/7/7$. Navtest means shift by at most 0.27 pp, preserving the ranking and the reference target's $+3.90$ pp native-relative gain. Appendix N reports the corresponding navhard controls. This does not establish scene-dependent command use or a bound for every seed or scene.

Table 14: **Navtest input controls.** EPDMS $\times 100$, mean $\pm$ sample SD across three seeds. Test-only keeps primary checkpoints; fully matched repeats training and checkpoint selection with constant context. The primary protocol uses recorded validation and test context.

Target	Primary	Test-only constant	Fully matched constant
Metric maximum	74.52 ± 0.84	74.79 ± 0.96	74.79 ± 0.95
Tie-uniform	74.25 ± 1.24	74.35 ± 0.96	74.35 ± 0.95
RethinkDrive (ours)	$\mathbf{87.65 \pm 0.12}$	$\mathbf{87.61 \pm 0.12}$	$\mathbf{87.61 \pm 0.12}$
Soft $\tau = 1$	82.99 ± 0.13	82.98 ± 0.13	82.97 ± 0.14

The measured 0.5193 M parameter total includes the instantiated auxiliary output heads, although their losses are inactive in the four main configurations; the broader comparison separately evaluates sub-score distillation. The main targets use the scalar logit in Eq. 3; sub-score-only distillation uses the separately described composed-score fusion.

H Bench2Drive evidence and limitations

H.1 Closed-loop execution and route coverage

A separate refinement experiment uses Trajectory-guided Control Prediction (TCP) (Wu et al., 2022) in Bench2Drive (Jia et al., 2024), executing actions in CARLA (Dosovitskiy et al., 2017). Its ten candidates are TCP's trajectory, six lateral offsets and three shortened variants. This changes the backbone and candidate construction; it does not repeat the four-target study.

The records contain 140 completed routes, 40 simulator runtime terminations, 13 blocked-agent terminations and one setup failure. Completed routes average DS 66.92, so completion alone

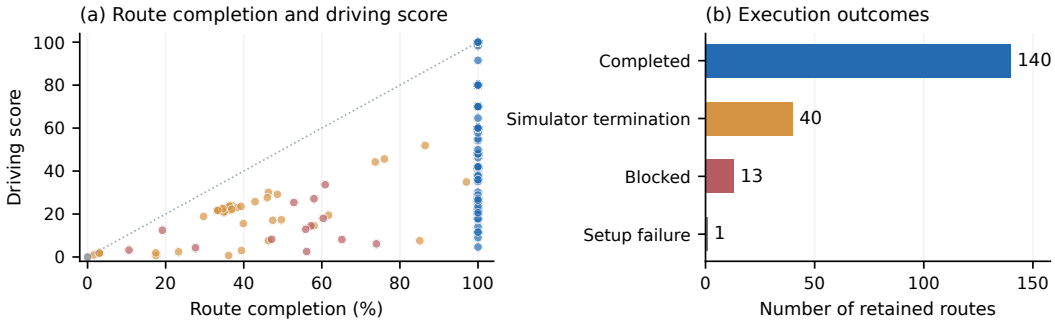

Figure 6: **Closed-loop completion does not imply a high driving score.** Each point in (a) is one of 194 retained route records; the diagonal is the no-penalty relation. Colors match the statuses in (b). Many completed routes incur penalties. Counts describe this refinement evaluation, not improvement over a matched TCP-only baseline. Overlapping points are not displaced.

Table 15: **Bench2Drive scores with explicit route coverage.** DS: driving score; RC: route completion; SR: routes with DS 100. The archived and retained sets differ; equal rounded DS does not reproduce the larger aggregate. Assigning zero to missing scores or records changes the denominator to all 220 planned routes.

Evidence and denominator	Routes	DS ↑	RC (%) ↑	SR (%) ↑
Archived scored aggregate	211	53.11	83.83	25.59
Retained route records	194	53.11	83.81	25.26
Archive; nine missing scores set to zero	220	50.94	—	—
Records; 26 missing records set to zero	220	46.83	—	—

overstates successful driving. A matched TCP-only evaluation is unavailable, and no refinement gain is inferred. The following subsections detail record provenance, missing routes and the separate development-route transfer results.

H.2 Record provenance and comparison scope

The archived status counts are 153 completed, 45 runtime terminations and 13 blocked; these statuses do not define success rate. The retained 194 records are a subset of the 211-route aggregate. Retained records use the last completed attempt when available, otherwise the last attempt, including partial scores from runtime terminations. The manifest identifies the missing routes. Table 16 reports only the retained records.

The TCP pool comprises KEEP, lateral offsets $\pm\{0.5, 1.0, 1.5\}$ m, and trajectories shortened by 5%, 10% and 15%. This construction does not ensure drivable-area compliance. No matched TCP-only run on this route set is available; the formerly quoted baseline 40.70 therefore cannot support a refinement increment. The earlier 14.3% ADE reduction and oracle-headroom claim are also excluded because prediction and reference horizons differed. Resampling does not establish a horizon-matched comparison.

Table 16: **Retained Bench2Drive route outcomes.** The 194 available records are a subset of the 211-route archived aggregate. Partial scores from terminated routes are included. The retained CSV supplies route identities and infraction counts.

Status	Routes	Mean DS
Completed	140	66.9
TickRuntime termination	40	18.9
Agent blocked	13	13.6
Agent setup failed	1	0.0

Corrected development-route transfer. The ten routes span Towns 03, 04, 05, 11, 12, 13 and 15. In this exploratory report, a successful route has DS 100; evaluator "Completed" status is a separate count. Nearest (80) and Nearest (80) + length each achieve one such route, correcting an earlier all-zero statement. These are earlier configurations and do not repeat the present four-target comparison. They do not establish that in-domain retraining is necessary under all possible transfer methods.

Table 17: **Earlier zero-shot dev10 results.** One run per row; Nearest (80) + length uses its second attempt because the first produced no routes after agent setup failure.

Configuration	DS 100	Mean DS	Mean RC	Completed / terminated
Frozen planner default	0/10	14.2	39.5	1/9
Nearest (80) (earlier configuration)	1/10	28.3	52.8	1/9
Nearest (80) + length (earlier configuration)	1/10	27.2	41.1	2/8
Refinement variant	0/10	26.9	63.9	4/6

I SELECTOR TIMING PROTOCOL

Table 18 measures the selector's forward pass with 800 candidates. Candidate generation and feature extraction are excluded; the reported latency therefore describes the added module rather than the complete driving pipeline.

Table 18: **Measured cost of the selector** (RTX PRO 6000, fp32, 800 candidates, 30 warm-up iterations, CUDA synchronisation in the timing loop, 200 repetitions).

	batch 1	batch 8
parameters (all trainable)	0.519 M	
added latency, mean / p95	0.56 / 0.57 ms	0.87 / 0.87 ms
added latency per scene	0.56 ms	0.11 ms

J PRIMARY BEHAVIORAL AUDIT

The following results use recorded context at evaluation and the same fixed candidate pool as Table 2. They supersede the earlier observed-acceleration audit for interpreting the primary comparison. Score differences are percentage points unless stated otherwise.

Table 19: Per-seed primary score contrasts. Seed numbers match the training runs. Navhard entries are official combined scores, retaining the evaluator's aggregation.

Quantity	Seed 1	Seed 2	Seed 3
navtest: RethinkDrive − native (pp)	$+4.04$	$+3.96$	$+3.80$
navtest: RethinkDrive − metric target (pp)	$+12.28$	$+13.47$	$+13.63$
navtest: metric target − native (pp)	-8.24	-9.50	-9.83
navhard: RethinkDrive combined score	0.2264	0.2198	0.2243
navhard: native combined score	0.2079	0.2079	0.2079

Primary navtest behavior. For seed 1, RethinkDrive improves $6{,}027$ scenes, ties on $2{,}832$, and worsens $3{,}287$ relative to native selection. Across seeds, its new zero-score cases include 207–234 DAC failures, 78–83 NC failures, 7–14 DDC failures and four TLC failures. There are 3–4 cases with multiple failed gates. Categories overlap, so their ranges are not summed. Transitions include positive fractional gates becoming zero; counting only $1 \to 0$ would miss such failures. A zero score does not identify which failure occurred without examining its gates.

Against the matched-seed metric-target selector, RethinkDrive introduces 156–192 new zero scores and recovers 756–967. Its mean sub-scores increase on every reported axis except EP, which decreases by 7.4 pp. Relative to native selection, the additional changes not shown in Table 4 are LK $+1.8$, DDC $+0.6$ and TTC -0.2 pp. Metric-target training reduces DDC by 1.7 pp relative to native. These are measured sub-score changes, not a causal attribution of the aggregate gain.

Table 20: Stage-specific navhard contrasts relative to native selection. Official combined scores use the evaluator's aggregation; they are not the sum or simple average of these stage differences. All rows retain the shared synthetic-frame sensor convention.

Selector	Stage	Δ EPDMS	Δ DAC	Δ HC	Δ EC
RethinkDrive	One	$+0.93$	-6.3	$+4.9$	$+38.8$
RethinkDrive	Two	$+5.57$	$+2.5$	$+20.1$	$+41.8$
Metric target	One	-17.5	—	—	—
Metric target	Two	-9.3	—	—	—

What the audit establishes. The identical EC-availability mask rules out differences in missing EC coverage between methods. It does not explain the learned temporal behavior. Higher EC is consistent with better agreement between successive predicted trajectories, but candidate-index changes alone cannot measure this: indices need not identify the same physical motion across frames. Appendix K applies this spatial and temporal alignment to quantify differences between successive selected plans.

K SCORE DECOMPOSITION AND SUCCESSIVE-PLAN CONSISTENCY

Let $G = \mathrm{NC}\,\mathrm{DAC}\,\mathrm{DDC}\,\mathrm{TLC}$ and $N = 5\mathrm{EP} + 5\mathrm{TTC} + 2\mathrm{LK} + 2\mathrm{HC}$. With adjacency indicator w,

$$E_{\text{without EC}} = \frac{GN}{14}, \qquad E_N = \frac{GN}{14 + 2w}, \qquad E_{\mathrm{EC}} = \frac{2wG\mathrm{EC}}{14 + 2w}, \quad E = E_N + E_{\mathrm{EC}}. \tag{9}$$

The EC-excluded score renormalizes weights and is not an additive contribution. Only E_N and E_{EC} add to the official score. The CSV sub-scores already include the evaluator's human-penalty filter; reconstructing from them matches the supplied primary navtest scene scores to maximum absolute error 3.4×10^{-16}. The 10,040 EC-available scene identifiers are identical across methods; 2,106 others use denominator 14. Temporal masks and paired differences are aligned by scene identifier before log-level resampling.

Table 21: **Paired score decomposition on navtest.** Differences in percentage points. Intervals resample 136 logs with fixed checkpoints after per-scene seed averaging; they do not include training randomness. Covering zero does not establish equivalence.

Contrast	Quantity	Difference	95% interval
Ours − native	Official EPDMS	+3.93	[+3.43, +4.44]
Ours − native	EC excluded, renormalized	−0.49	[−1.01, +0.03]
Ours − native	Original-denominator E_N	−0.42	[−0.89, +0.04]
Ours − native	Original-denominator E_{EC}	+4.36	[+4.14, +4.56]
Ours − native	EC on adjacent-frame scenes	+45.07	[+43.03, +47.09]
Ours − metric	EC excluded, renormalized	+7.53	[+6.55, +8.58]
Tie-uniform − metric	EC on adjacent-frame scenes	−1.19	[−2.09, −0.29]

Dynamic-state comparison. The scorer simulates each selected eight-waypoint plan with its bicycle model and tracker, at 0.1 s resolution. Consecutive original frames are 0.5 s apart. The overlapping 35 states compare the current plan with the previous plan shifted by five steps in a common global frame. The scorer's smoothed acceleration, jerk, yaw-rate and yaw-acceleration features produce RMS differences. EC passes when all stay below their respective thresholds: 0.7 m/s^2, 0.5 m/s^3, 0.1 rad/s and 0.1 rad/s^2. Recomputed threshold decisions match exported EC on all 200,800 evaluated method-pair rows. Recorded human futures pass through the same simulator as a descriptive reference.

Table 22: **Differences between successive simulated plans.** Mean RMS differences over the same 10,040 pairs; learned rows pool three seeds. Lower RMS means more consistent plans, not necessarily safer plans. The human row is a descriptive reference.

	EC pass	Acceleration m/s^2	Jerk m/s^3	Yaw rate rad/s	Yaw accel. rad/s^2
Recorded human	0.901	0.121	0.219	0.022	0.026
Native planner	0.340	0.454	0.721	0.041	0.054
Metric target	0.143	0.729	1.156	0.055	0.075
Tie-uniform	0.132	0.674	1.068	0.057	0.074
Metric + reference tie-break	0.267	0.547	0.845	0.046	0.060
RethinkDrive (ours)	**0.791**	**0.235**	**0.355**	**0.032**	**0.039**
Soft $\tau = 0.3$	0.366	0.439	0.676	0.041	0.053
Soft $\tau = 1$	0.296	0.481	0.746	0.042	0.056

These measurements quantify the output differences that EC thresholds. They do not prove that index-dependent supervision causes those differences. Tie-uniform retains low EC despite removing the first-index preference. Recorded-speed stratification is descriptive, and does not hold the selected speed or future behavior fixed.

Two-stage aggregation. Navhard exports permit stage-specific decomposition, but not reconstruction of a combined EC-excluded score: the Gaussian endpoint weights are absent. Uniform weighting does not reproduce the official combined result. We therefore report stages separately and retain the scorer's official combined EPDMS, without adding or averaging the stage differences.

L TIE RESOLUTION AND TARGET CONCENTRATION

Evaluation-aligned tie diagnosis on navtest. Every candidate in all 12,146 stored pools is scored using the single-stage evaluation criteria, with eight poses, no endpoint padding and no two-frame EC. Saturation occurs in 74.4% of scenes and tied maxima in 77.4%; the conditional median tied-set size is 81. All executed trajectories used in this audit match a stored candidate. We associate these scene properties with the primary selectors' official scores, averaged over three seeds.

Table 23: **The metric-target deficit is not confined to evaluated ties.** Contrasts are in pp. Δ_{R-M} compares RethinkDrive with metric-target training; Δ_{R-N} compares it with native. Without-EC renormalizes the score. Raw EC is a sub-score on available adjacent frames, not an additive contribution.

Tied-set size	Scenes	Δ_{R-M}	Δ_{R-N}	Without-EC Δ_{R-N}	Raw EC Δ_{R-N}
1 (unique)	2,749	+13.42	+2.97	−1.71	+47.95
2–10	1,626	+10.36	+3.66	−0.94	+46.77
11–50	2,262	+12.09	+3.66	−1.42	+49.62
51–200	2,566	+13.48	+3.70	−0.60	+43.16
201–800	2,943	+14.86	+5.40	+1.71	+39.45

Across all tied scenes, $\Delta_{R-M} = 13.04$ pp, compared with 13.42 pp for unique maxima. The Spearman correlation between tied-set size and this contrast is 0.042. These associations do not show that saturation causes the learning deficit. In particular, a unique maximum under this evaluated score need not be unique under the cached training score. The matched-subset training intervention in Appendix M tests aligned labels separately. Neither result identifies saturation as the cause of the remaining learned-performance deficit.

Reference tie-break within metric maxima. The control selects the smallest-displacement candidate inside the cached tied set, leaving all other primary training settings unchanged. It changes about 80% of the metric labels and agrees with unrestricted reference labels in 37.2% of training scenes. Three seeds reach 79.13 ± 1.59 on navtest and 15.03 ± 0.81 on navhard. The navtest improvement over first-index supervision is 4.61 pp, conditional log-bootstrap interval $[4.12, 5.13]$. The remaining deficit to unrestricted reference supervision is 8.52 pp. This establishes a benefit from the tested tie-resolution rule, not the fraction of the full deficit caused by saturation. Uniform mass over the tied set also removes a first-index preference but does not recover the same gain.

Temperature changes behavior as well as score. The median target entropy falls from 6.15 nats at $\tau = 1$ to 4.77 at $\tau = 0.1$; nearest-candidate mass rises from 0.4% to 2.7%. Effective support decreases from about 470 to 118 of 800 candidates. The lower-temperature target remains soft. Table 24 reports the completed three-seed contrasts. Its navtest shortfall to the hard target is accompanied by better EC-excluded scores; both lower temperatures exceed the hard target on navhard. These results delimit the scope of the hard target's advantage without selecting a temperature on test performance.

Table 24: **Soft-minus-hard contrasts.** All configurations have three training seeds. Navtest intervals resample logs, conditional on the retained checkpoints. Navhard uses its official combined score and has no EC-excluded combined reconstruction.

Temperature	Navtest EPDMS	Navtest without EC	Navhard EPDMS
$\tau = 0.3$	-3.47 $[-3.95, -3.02]$	$+0.70$ $[0.22, 1.19]$	+1.96
$\tau = 0.1$	-0.66 $[-0.97, -0.35]$	$+0.85$ $[0.57, 1.14]$	+1.07

M METRIC-ALIGNMENT PILOT AND SECOND-PLANNER EVALUATION

M.1 MATCHED-SUBSET LABEL ALIGNMENT

The pilot freezes a subset of 20,279 training scenes for which all 800 candidates have been re-scored. All three targets use exactly this subset, the same predicted input features, safety penalty, fusion rule, eight-epoch budget and three seeds. The aligned metric uses eight 0.5 s poses without endpoint padding and the single-stage evaluation score without two-frame EC, with denominator 14. It supplies a first-maximum label; 84.4% of pilot scenes have tied maxima and 81.6% saturate. The controls rebuild the cached-metric and nearest-reference targets on the same subset. None of these rows is compared to a full-data target as an isolated label intervention.

Training context is constant; validation and test use recorded context, matching the primary protocol. Checkpoints are selected on the complete token-held-out validation set using cached PDMS, with unchanged patience. This preserves comparability across the three pilot targets but does not align the checkpoint criterion or the omitted temporal term. The aligned target selects epoch 1 for all three seeds. Table 25 reports official navtest evaluation on all 12,146 scenes. The corresponding navhard results and fully matched context controls are in Appendix N.

Table 25: **Label alignment on the same training subset.** Official navtest scores $\times 100$; mean $\pm$ sample SD of three seeds. The three learned rows use 20,279 training scenes. Native is a fixed test-time reference, not another pilot-trained row.

Target	EPDMS ↑	Without EC	EC	Δ native (pp)
Planner, native selection	83.71	89.62	33.99	—
Metric maximum	75.15 ± 0.66	82.34	14.18	−8.57
Aligned metric maximum	78.82 ± 1.76	85.69	20.95	−4.89
RethinkDrive (ours)	$\mathbf{87.33 \pm 0.16}$	**89.33**	**73.74**	**+3.62**

Aligned versus cached labels improve the official score by 3.67 pp, conditional 95% log-bootstrap interval $[3.26, 4.10]$, and the EC-excluded score by 3.35 pp, interval $[2.93, 3.80]$. Aligned labels remain 8.51 pp below reference supervision, interval $[-9.07, -7.99]$, and 4.89 pp below native, interval $[-5.38, -4.44]$. Thus correcting the tested label-definition and pose conventions helps, but does not close the observed gap. These contrasts do not isolate the remaining cause, prove a general disadvantage of metric supervision, or cover an optimized full-data alignment recipe. The tie-break and alignment interventions use different training populations; their gains cannot be added or interpreted as fractions of a common causal decomposition.

M.2 DRIVESUPRIM PROTOCOL

The second planner is a frozen DriveSuprim ResNet-34 checkpoint (Yao et al., 2026). Its vocabulary has 8,192 trajectories, each with 40 poses at 0.1 s intervals. Indices $4, 9, \ldots, 39$ form eight 0.5 s poses for the shared feature and evaluation conventions. A fixed shortlist of 256 candidates is reconstructed from the planner's coarse composite before any selector target is constructed. Every learned target uses that same shortlist. The native planner's output lies in it in 98.5% of navtest scenes; for the remaining 1.5%, the added selector cannot reproduce native selection.

Let p_i^{imi} be the vocabulary softmax probability and $u_{i,r}$ the sigmoid of predicted metric-head logit r. The coarse composite is

$$c_i = 0.02 \log p_i^{\mathrm{imi}} + 0.1 \log u_{i,\mathrm{TLC}} + 0.5 \log u_{i,\mathrm{NC}} + 0.5 \log u_{i,\mathrm{DAC}} + 0.3 \log u_{i,\mathrm{DDC}} + 6\,(5u_{i,\mathrm{TTC}} + 5u_{i,\mathrm{EP}} + 2u_{i,\mathrm{LK}})/12. \tag{10}$$

Within the shortlist, the frozen fusion score is min–max normalized: $f_i = (c_i - \min_j c_j)/(\max_j c_j - \min_j c_j)$, with a numerical floor on the denominator. This scale adapter is common to all four learned targets. The nine base-input slots are f_i, p_i^{imi} and the seven predicted metric heads. Geometry supplies the same 16 features as in Appendix G; command and ego state are constant during training, validation and test. Changing the base-input width gives a 0.520 M-parameter selector. Inference latency is not measured for this adapter.

Metric-maximum and tie-uniform targets use the vocabulary's cached metric scores; reference and soft targets use displacement on the shortlisted candidates, with $\tau = 1$ for the soft target. The expected NC/DAC/TTC penalty and sigmoid score fusion are shared. Each target has three seeds under the common optimizer and checkpoint rule. This transfers the comparison design while adapting candidate support and input scores; it does not hold the generator, feature calibration or native-selection mechanism fixed across planners.

Table 26: **DriveSuprim with a frozen top-256 shortlist.** Navtest EPDMS $\times 100$; learned rows report three-seed sample SD. Conditional paired-log intervals use seed-averaged scene differences. All four learned-target intervals include zero.

Target	EPDMS ↑	Δ native	95% interval	EC
Planner, native selection	86.54	—	—	77.33
Planner, coarse score only	86.82	+0.29	$[-0.24, 0.83]$	76.33
Metric maximum	86.70 ± 0.06	+0.16	$[-0.33, 0.69]$	76.57
Tie-uniform	86.38 ± 0.15	−0.16	$[-0.65, 0.35]$	75.39
RethinkDrive (ours)	$\mathbf{86.83 \pm 0.16}$	**+0.29**	$[\mathbf{-0.18, 0.81}]$	**79.54**
Soft reference ($\tau = 1$)	86.92 ± 0.10	+0.38	$[-0.11, 0.91]$	78.38

All native-relative intervals include zero. The reference target's mean change is $+0.29$ pp, interval $[-0.18, 0.81]$, compared with $+3.93$ pp for DDv2 in the primary experiment. This is absence of a statistically resolved gain under the tested protocol, not an equivalence result. A formal planner-by-target interaction test is not supplied. DriveSuprim's coarse-only baseline scores 86.823 and the reference selector 86.828; this comparison also does not demonstrate added value from learning. Native EC is 77.3, versus 34.0 for DDv2. These observations motivate a hypothesis about the room for improvement, but do not establish why the target effect differs. The DriveSuprim experiment evaluates navtest only and does not establish transfer to navhard or closed-loop driving.

Reproducibility and numerical precision. The supplement includes the target builders, selector and adapter code, launch scripts, resolved configurations and result manifests. Per-scene score exports identify the log, validity and every score component. Aggregation reconstructs the official score and EC-excluded score, averages seeds before paired log resampling, and reports seed SD separately. The supplied tables are rounded from scene-level aggregates; they are not reconstructed by subtracting already rounded display values. Stored checkpoint and prediction hashes identify the runs. Six primary DDv2 selector checkpoints and their inference derivatives are supplied separately. Appendix O supplies six more log-disjoint checkpoints; planner weights, caches and other selectors remain external.

Complete primary sub-scores. Table 27 retains history comfort, collision and lane-keeping criteria alongside the compact main-text comparison. Individual sub-score differences cannot be summed to recover the aggregate score because of gates and varying EC weights.

Table 27: **Comfort improves while progress and compliance decrease.** Navtest scores $\times 100$; learned columns average three seeds. EC uses the same 10,040 adjacent-frame scenes; other rows use all 12,146. Without-EC renormalizes the remaining weights. Individual sub-scores are not additive EPDMS contributions.

Quantity	Native	Metric	Tie-uniform	**Ours**	Soft $\tau = 1$
EPDMS	83.7	74.5	74.3	**87.6**	83.0
Without EC, renormalized	89.6	81.6	81.4	**89.1**	89.3
Extended comfort (EC)	34.0	14.3	13.2	**79.1**	29.6
History comfort (HC)	91.8	70.7	75.5	**98.0**	91.6
Ego progress (EP)	91.2	95.2	95.5	**87.8**	89.5
Drivable area (DAC)	97.8	95.0	95.1	**96.7**	97.8
No at-fault collision (NC)	98.3	95.6	95.2	**98.2**	98.5
Lane keeping (LK)	94.2	92.7	91.9	**96.1**	94.1
Time to collision (TTC)	97.8	94.3	93.7	**97.6**	98.0

N ADDITIONAL NAVHARD CONTROLS AND REPRODUCIBILITY COVERAGE

Additional navhard controls. Table 28 evaluates the matched-subset alignment pilot and the fully matched context control using official two-stage outputs. All 21 learned runs and the native baseline contain 450 valid first-stage and 5,462 valid second-stage scenes, without a single-stage fallback. They use the same candidate cache and the same empty-LiDAR convention as the primary navhard comparison.

Table 28: **Additional navhard controls.** Official combined EPDMS $\times 100$; mean $\pm$ sample SD over three seeds. Native selection is fixed. Comparisons between pilot and full-data rows do not isolate a target intervention.

Protocol	Target	Combined EPDMS ↑	Δ native (pp)
Native planner	Native selection	20.79	—
20,279-scene pilot	Metric maximum	11.58 ± 0.49	-9.20
	Aligned metric maximum	15.43 ± 2.63	-5.36
	RethinkDrive (ours)	22.85 ± 1.15	$+2.07$
Constant context	Metric maximum	11.17 ± 0.76	-9.62
	Tie-uniform	10.64 ± 0.38	-10.15
	RethinkDrive (ours)	22.50 ± 0.34	$+1.71$
	Soft reference ($\tau = 1$)	22.25 ± 0.14	$+1.47$

Within the pilot, aligned labels improve on cached-metric labels by 3.84 pp but remain 7.43 pp below reference labels. This is a partial label-alignment result: selection still uses cached validation PDMS, and aligned targets omit two-frame EC. With constant context throughout training, validation and test, the reference selector scores 22.50, 1.71 pp above native. This input control retains the validation dependence described in Appendix B.

What was verified. The released verification script reads the combined summary written by the official scorer, checks the associated scene counts and output hashes, and recomputes the three-seed means and sample SDs. It does not independently reconstruct the two-stage combined value: Gaussian endpoint weights are absent from the exported scene rows. Navtest scene-level means can be re-aggregated directly.

Executable inference and external assets. The checkpoint archive provides the primary DDv2 reference and metric selectors for all three seeds: six original retained checkpoints and six inference derivatives. The packaged execution record reports identical original/derivative picks on three smoke scenes and 500 validation scenes per checkpoint, with zero fused-score difference. A separate six-checkpoint smoke record passes all 18 checkpoint–scene comparisons. Inputs are candidate geometry, predicted planner scores and recorded context; neither the human trajectory nor simulator labels are inference inputs.

The supplement includes exact split lists, environment/configuration records, path preparation and score-verification scripts. These checks cover inference and aggregation, not full retraining, candidate generation or simulator re-execution. Remaining external assets include planner weights, sensor data, candidate and metric caches, and unreleased selector checkpoints. Primary per-epoch checkpoints were not retained and the original split-builder script is absent; exact scene lists fix that split's membership. The additional log-disjoint experiment releases its split builder and six selected checkpoints, with hashes for retained per-epoch files; those epoch files are not all included. The large DriveSuprim training-label file and original prediction dumps still lack complete hashes. Appendix O distinguishes the additional evidence from this primary archive.

O ADDITIONAL CONTROLS FOR MODEL SELECTION AND LEARNED REFINEMENT

These experiments were designed after the primary test results were available. Their split rule, candidates, seeds, optimizer, stopping rule and parameter grids were fixed before the additional results. They are follow-up controls, not a preregistered study. Primary results retain their original protocol; Table 3 identifies the new log-disjoint runs separately.

O.1 LOG-DISJOINT RETRAINING

The population combines 87,392 original training and 5,278 validation scenes: 92,670 scenes from 1,191 recording logs. Each log receives a salted SHA-256 key; the first $\lfloor 0.1 \times 1{,}191 \rfloor = 119$ logs in sorted order form validation. The salt, scene-to-log mapping and assignment script are released. This gives 85,077 training scenes in 1,072 logs and 7,593 validation scenes in 119 logs. Token and log intersections are empty. No scene is dropped for missing candidates or targets. Official test logs are disjoint from both new subsets. Target indices were re-derived on 200 scenes from each original source partition, agreeing in all 400 cases.

Reference, metric maximum and soft temperatures $\{1, 0.3, 0.1\}$ are each retrained with three random initializations. Three further runs remove the reference selector's safety penalty. All 18 runs use the common eight-epoch maximum, patience three, AdamW settings and fixed pools. Context is constant in training, validation and test. Checkpoints maximize cached validation PDMS of the fused pick. Temperature selection maximizes the three-seed mean of that validation score, with ties favoring larger τ. It selects $\tau = 1$ again: validation means are 0.9444, 0.9402 and 0.9336 for $\tau = 1, 0.3, 0.1$, respectively.

Table 29: **Uncertainty under log-disjoint model selection.** Navtest contrasts against native selection, in pp. Intervals resample 136 logs of seed-averaged scene contrasts with 10,000 draws, conditional on models and candidates. Seed SD is reported separately in Table 3.

Target / configuration	Δ native	95% interval
Metric maximum	-9.55	$[-10.35, -8.74]$
RethinkDrive (ours)	$\mathbf{+3.93}$	$[\mathbf{3.39}, \mathbf{4.46}]$
Soft $\tau = 1$ (validation-selected)	-0.87	$[-1.22, -0.54]$
Soft $\tau = 0.3$	$+0.44$	$[0.04, 0.85]$
Soft $\tau = 0.1$	$+3.52$	$[3.04, 3.99]$
RethinkDrive, no penalty	$+3.85$	$[3.30, 4.40]$
RethinkDrive, selected fusion $\lambda = 0.3$	$+4.21$	$[3.73, 4.68]$

The reference–metric contrast remains 13.47 pp on navtest, interval $[12.48, 14.49]$. Reference navhard scores are 22.977, 22.683 and 22.969, versus metric scores 10.444, 11.448 and 10.975. All additional navhard runs complete 450 first-stage and 5,462 second-stage observations. Combined values are read from the official scorer summary; absent Gaussian endpoint weights prevent independent reconstruction. These results support robustness to this split change. They do not estimate a causal leakage effect, because training membership and size also change. Cached validation remains imperfect: its selected temperature is not best on either test split.

Evidence coverage. The release includes 21 new reference/metric/soft/penalty/fusion navtest scorer outputs and 18 new navhard outputs, split lists, configurations, training logs and six selected reference/metric checkpoints. A per-epoch hash inventory identifies retained files; not every epoch checkpoint is packaged. The original dependent split remains the primary protocol, explicitly distinguished from this follow-up.

O.2 Learned ranking and score fusion

For primary DDv2 and second-planner checkpoints, we compare selector-only execution $\arg\max_i z_i$ with fused execution $\arg\max_i [f_i + \sigma(z_i)]$. This changes execution without retraining or reselecting a checkpoint. Original checkpoint selection used fused scores; the comparison diagnoses those models, rather than optimizing an unfused model.

Table 30: **Target effects depend on fusion and planner.** Navtest EPDMS $\times 100$. Selector-only columns show three-seed mean $\pm$ SD. Fused means use the same checkpoints under the primary DDv2 or matched DriveSuprim protocol. Δ compares selector-only with native. Native/coarse scores are $83.71/81.69$ for DDv2 and $86.54/86.82$ for DriveSuprim.

Planner	Target	Selector only	Fused	Δ native
DDv2	Metric maximum	62.13 ± 2.63	74.52	-21.59
	Tie-uniform	61.39 ± 0.28	74.25	-22.32
	RethinkDrive (ours)	$\mathbf{86.44 \pm 0.06}$	**87.65**	$\mathbf{+2.73}$
	Soft $\tau = 1$	87.18 ± 0.03	82.99	$+3.47$
DriveSuprim	Metric maximum	80.44 ± 1.51	86.70	-6.10
	Tie-uniform	72.27 ± 0.91	86.38	-14.27
	RethinkDrive (ours)	$\mathbf{81.83 \pm 0.59}$	**86.83**	$\mathbf{-4.70}$
	Soft $\tau = 1$	83.37 ± 0.23	86.92	-3.17

DDv2 reference ranking alone exceeds native by 2.73 pp, interval $[2.08, 3.36]$. Fusion adds 1.21 pp, while the metric-trained ranking alone performs poorly. Soft supervision differs: removing fusion improves its score by 4.19 pp and reverses its ordering against the unfused hard target. The primary hard–soft contrast is conditional on fusion. Equal coefficients do not imply equally calibrated learned scores; no hard-target superiority independent of execution is supported.

Why the DriveSuprim gain does not emerge in this protocol. The reference selector alone increases EC from native 77.33 to 80.98, but reduces DAC from 96.51 to 91.91 and the renormalized non-EC score from 88.14 to 82.88. Its total is consequently 81.83, below native 86.54. Fusion raises DAC to 97.07 and total score to 86.83, near coarse-only 86.82, while reference EC remains 79.54. This identifies a behavioral trade-off and recovery under fusion, not an absence of behavioral change: 45.7% of coarse choices change, averaged over three seeds. Across the four target means, the selector-only range of 11.10 pp contracts to 0.54 pp after fusion. These ranges are descriptive, not interaction-test statistics. Candidate support and score normalization differ from DDv2 (Appendix M); neither their causal roles nor target quality and available improvement are separately identified. High native EC alone does not prove that improvement is impossible.

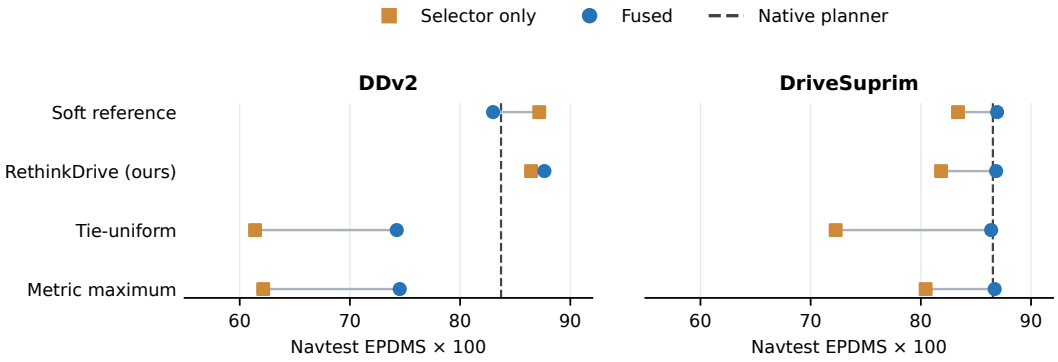

Figure 7: **Fusion changes the effect of a target.** Three-seed means from Table 30; segments connect the same checkpoints under two execution rules. On DDv2, fusion helps the hard reference target but hurts soft $\tau = 1$. On DriveSuprim, fusion brings all four means near native selection; its intervals include zero (Table 26). Segments are not uncertainty intervals.

O.3 TRAINING-FREE RULES AND VALIDATION-SELECTED FUSION

Two rules score candidates as $f_i - \lambda p_i$. For length, p_i is arc length from the ego origin through all eight waypoints, min–max normalized across candidates, or zero if lengths coincide. This bounded preference differs from unconditional shortest-path selection. For consistency, p_i is normalized RMS position difference from the plan this rule selected at the preceding observation of the same log. Recorded ego poses transform that plan to the current ego frame; the time shift uses observed frame separation. Comparison uses the overlapping horizon without extrapolation. Scenes are processed chronologically. With no preceding observation within 0.55 s, the rule uses coarse scores alone. Future human trajectories and cached simulator scores are unavailable at inference. Additional plan history makes this a different-input comparator.

Rule and reference-fusion weights use $\{0, 0.1, 0.3, 1\}$. Selection maximizes official one-stage validation EPDMS; reference fusion averages three seeds and ties favor the smaller weight. Length selects 0, consistency 0.3 and reference fusion 0.3. Zero weight returns coarse selection exactly.

Table 31: **Validation-selected rules and fusion sensitivity.** EPDMS $\times 100$. Bold marks each validation-selected weight. Reference rows average three seeds; rules are deterministic. Reference at test weight 0.1 was not scored. Additional validation digits distinguish close maxima.

Split	Selection	$\lambda = 0$	0.1	0.3	1
Validation	Length penalty	**85.846**	84.965	81.636	75.187
	Previous-plan consistency	85.846	86.063	**86.089**	85.960
	RethinkDrive fusion	85.846	91.586	**91.934**	91.893
navtest	Length penalty	**81.69**	81.01	77.79	72.96
	Previous-plan consistency	81.69	81.79	**81.35**	80.31
	RethinkDrive fusion	81.69	—	**87.93**	87.64

The selected consistency rule is 0.34 pp below coarse, interval $[-0.70, 0.02]$. EC rises from 23.6 to 26.7, less than the reference selector's increase to 79.0. These rules do not reproduce the learned gain; this does not establish that all training-free refinements fail. The previous-plan rule is not evaluated on navhard: a corresponding history is undefined for synthetic stages. Selected reference fusion is also evaluated only on navtest; its navhard entry is unreported.

Selected reference fusion improves over $\lambda = 1$ by 0.29 pp, interval $[0.12, 0.45]$. Its renormalized without-EC contrast against native is -0.03 pp, interval $[-0.52, 0.44]$; this is not equivalence. Progress still falls by about 3.5 points and DAC by 0.6. New-zero/recovered counts average $241/234$, compared with $320/241$ at $\lambda = 1$. The aggregate gain retains behavioral costs.

Safety-penalty ablation. Removing the penalty changes log-disjoint navtest EPDMS by -0.07 pp, interval $[-0.17, 0.03]$; the without-EC contrast is -0.12 pp, interval $[-0.22, -0.01]$. Navhard scores are 20.99, 23.25 and 23.54: 22.59 ± 1.40 versus 22.88 ± 0.17 with the penalty. The larger spread includes an early-selected first-seed checkpoint. No independent aggregate contribution is established; the penalty remains part of the shared primary recipe.

O.4 PRIMARY COMPONENT VARIANTS AND EVIDENCE SCOPE

Table 32: **Component changes have smaller effects than the primary target contrast.** EPDMS $\times 100$; three-run rows show mean $\pm$ SD. Native scores are $83.7/20.8$. DPO and hybrid use the first-80 plus length target; DPO and policy gradient omit the safety penalty, and sub-score distillation changes fusion. These are broader controls, not target-only ablations. Bold identifies our primary configuration.

Configuration	Runs	navtest ↑	navhard ↑
Random candidate	—	52.30	8.33
Shortest trajectory	—	42.27	15.87
Longest trajectory	—	7.83	0.57
Policy gradient with metric reward	1	81.27	17.56
Sub-score distillation	1	80.36	19.56
RethinkDrive (ours): nearest of 800	3	$\mathbf{87.65 \pm 0.12}$	$\mathbf{22.35 \pm 0.34}$
Nearest target restricted to first 80	3	87.63 ± 0.13	22.84 ± 0.88
First-80 target + length matching	3	87.56 ± 0.05	22.54 ± 0.78
DPO-style, first-80 + length	1	87.58	23.26
Hybrid, first-80 + length	1	87.56	23.05
Soft $\tau = 1$ + additional ranking loss	3	82.39 ± 0.03	21.56 ± 0.09

Restricting target support to the first 80 changes 73.9% of training labels; length matching changes 69.9% relative to that target. Reference-based variants choose different trajectories (17–50% agreement in the audited scenes), despite similar navtest means. DPO-style and hybrid variants also change objectives; policy-gradient and sub-score rows change other recipe elements. They are broader component controls, not independent target-only replications. Table 3 separately tests model-selection sensitivity.

Primary soft $\tau = 0.1$ is 0.66 pp below the hard target on navtest, but its without-EC score is 0.85 pp higher, interval $[0.57, 1.14]$. Relative performance depends on split, components and fusion. Complete primary and log-disjoint temperature sweeps are reported without replacing the validation-selected temperature after test inspection.

Calibration scope. Adding the same constant to every logit leaves the softmax-based loss in Eq. 2 unchanged, but can change the fused choice in Eq. 3. Equal fusion weights therefore do not ensure equal calibration across targets. The unfused and weight-selection controls expose this dependence; they do not establish an optimal fusion rule or isolate the effect of a common logit shift.

Released records and verification limits. Additional summaries identify target, protocol, seed, selected epoch and checkpoint, prediction and scorer hashes. They contain 91 completed scorer records. Three unexecuted zero-fusion validation rows are marked separately and use the identical coarse baseline. Released navtest outputs permit score and EC-decomposition re-aggregation; navhard verification reads the combined summary. Retraining, candidate generation and simulation require the external assets in Appendix N.

The additional archives include all 85 new scorer outputs: 51 navtest, 16 validation and 18 navhard runs; six baseline records link to the preceding evidence. They also include per-scene selection indices, rule-selection grids and six additional reference/metric checkpoints. Other target and per-epoch weights are not all bundled. Execution and distributed-code hashes are distinguished in the release notes. Intervals quantify conditional scene/log uncertainty, not variability from regenerating pools or repeating the whole study. These results do not establish real-road safety.

SUPPLEMENTARY REFERENCES

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank using Gradient Descent. In *ICML*, 2005.

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

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

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

Sergio Casas, Abbas Sadat, and Raquel Urtasun. MP3: A Unified Model To Map, Perceive, Predict and Plan. In *CVPR*, 2021.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. MultiPath: Multiple Probabilistic Anchor Trajectory Hypotheses for Behavior Prediction. In *CoRL*, 2020.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-Driven Non-Reactive Autonomous Vehicle Simulation and Benchmarking. In *NeurIPS*, 2024.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *CoRL*, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. ST-P3: End-to-end Vision-based Autonomous Driving via Spatial-Temporal Feature Learning. In *ECCV*, 2022.

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

Zhiyu Huang, Haochen Liu, Jingda Wu, and Chen Lv. Differentiable Integrated Motion Prediction and Planning With Learnable Cost Function for Autonomous Driving. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think Twice before Driving: Towards Scalable Decoders for End-to-End Autonomous Driving. In *CVPR*, 2023.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-To-End Autonomous Driving. In *NeurIPS*, 2024.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In *ICCV*, 2023.

Bo Jiang, Shaoyu Chen, Hao Gao, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. VADv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning. In *ICLR*, 2026.

Jungho Kim, Jiyong Oh, Seunghoon Yu, Hongjae Shin, Donghyuk Kwak, and Jun Won Choi. SafeDrive: Fine-Grained Safety Reasoning for End-to-End Driving in a Sparse World. In *CVPR*, 2026. arXiv:2602.18887.

Kailin Li, Zhenxin Li, Shiyi Lan, Yuan Xie, Zhizhong Zhang, Jiayi Liu, Zuxuan Wu, Zhiding Yu, and Jose M. Alvarez. Hydra-MDP++: Advancing end-to-end driving via expert-guided hydra-distillation. *arXiv preprint arXiv:2503.12820*, 2025a.

Yaguang Li, Jiaru Zhang, Chuheng Wei, Can Cui, and Ziran Wang. Designing Versatile Samples for Learned Trajectory Scoring. *arXiv preprint arXiv:2609.01799*, 2026.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing End-to-End Autonomous Driving with Latent World Model. In *ICLR*, 2025b.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In *ECCV*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

Tung Phan-Minh, Elena Corina Grigore, Freddy A. Boulton, Oscar Beijbom, and Eric M. Wolff. CoverNet: Multimodal Behavior Prediction Using Trajectory Sets. In *CVPR*, 2020.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. SparseDrive: End-to-End Autonomous Driving via Sparse Scene Representation. In *ICRA*, 2025.

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. SparseDriveV2: Scoring is All You Need for End-to-End Autonomous Driving. *arXiv preprint arXiv:2603.29163*, 2026.

Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided Control Prediction for End-to-end Autonomous Driving: A Simple yet Strong Baseline. In *NeurIPS*, 2022.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise Approach to Learning to Rank: Theory and Algorithm. In *ICML*, 2008.

Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M. Alvarez, and Zuxuan Wu. DriveSuprim: Towards Precise Trajectory Selection for End-to-End Planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2506.06659.

Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-To-End Interpretable Neural Motion Planner. In *CVPR*, 2019.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. DiffusionDriveV2: Reinforcement Learning-Constrained Truncated Diffusion Modeling in End-to-End Autonomous Driving. *arXiv preprint arXiv:2512.07745*, 2025.