# When an Evaluation Rule Writes Training Labels: Measuring Human-Reference Forgiveness in NAVSIM

**Anonymous authors**
Paper under double-blind review

## Abstract

When the human reference scores zero on a metric, the released GTRS-Dense label generator for NAVSIM marks every candidate trajectory in the scene as passing it. NAVSIM's authors introduced this human-reference forgiveness to avoid penalizing contextually justified maneuvers when scoring one trajectory, and warned that it could overlook important failures. In label generation it sets a whole column of 16,384 candidate targets to passing. To measure the consequences for supervision, we re-run the generator with the overwrite disabled and compare the pre-overwrite targets with the released labels on all 103,288 navtrain scenes. The rule erases a candidate distinction that the training loss reads on 11,237 of them (10.8793%). Firing usually changes most of a column: lane keeping carries 9,982 of the 13,042 forgiven loss columns, and its median forgiven column had 14,391 of 16,384 candidates failing before the overwrite. On held-out navtest scenes forgiven on lane keeping, the released lane-keeping head's median AUC against the pre-overwrite outcome is 0.7095; on unforgiven scenes matched on failing-candidate count it is 0.9807. For the Hydra-MDP checkpoint released with GTRS, whose configuration takes the same label file, the two values are 0.6627 and 0.9761. Continuing the released GTRS-Dense checkpoint for 300 optimizer steps with three paired seeds, we observe the forgiven-scene AUC 0.1086–0.1251 higher with pre-overwrite than with published targets, and a narrower gap between matched groups, still above zero. Scoring with forgiveness disabled, we observe lane keeping higher by 2.478–3.524 points on navtest scenes forgiven on any of five loss metrics, with lower adjacent-frame plan consistency. Both changes are larger there than on the rest. EPDMS, scored the same way, does not separate the two target sets.

## 1 Introduction

NAVSIM's authors introduced human-reference filtering to avoid penalizing contextually justified maneuvers and warned that it could overlook important failures and edge cases (Cao et al., 2025, Section 3.1; main-text Limitations, "Human Flag Filtering"). They called for further study of that trade-off, which concerns evaluation reliability, without measuring its effect on training labels; we measure that effect for the released GTRS-Dense trajectory labels. The measurement has two parts: how often the rule fires, and how many of the 16,384 candidate targets in a column it changes when it does, which on this split is most of them.

The same rule serves a different consumer when a scorer generates training targets. GTRS, whose dense scorer builds on Hydra-MDP (Li et al., 2024, Sections 2.3–2.4), learns component scores for candidate trajectories and uses their predictions to select a plan (Li et al., 2025, Sections 2.2–2.3 and 3). Its label generator replaces a component by passing labels throughout a scene's vocabulary when the human reference fails that component. An overwrite can change a target level without removing a target distinction: a column that already assigns every candidate the same failing value had no candidate ordering to erase. We therefore distinguish a human-reference zero, an overwrite that changes a value, and an overwrite of a previously mixed column. The last predicate is evaluated after the training consumer's target mapping and restricted to the columns its loss reads.



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.

We also measure whether vocabulary dropout would retain the pre-overwrite distinction and compare distinct label rows before and after overwrite within each token. We test the earlier split criterion, a v1 screen of navtrain on the human's PDM score (PDMS) rather than the rule measured here, with a separate v1 human census in Appendix B, and compare the top-ranked candidate sets of the two label files under a stated rule in Appendix D.

**Contributions.** We recover the pre-overwrite side of the released label file by re-running its generator with the forgiveness assignment disabled, and check that reconstruction against the published file on every column the overwrite did not touch (Section 4). That pair of arrays carries a per-column magnitude, which the single scored trajectory per agent of Wei et al. (2026) cannot (Section 2), and separates the human-zero trigger from a changed target level and from an erased distinction, after the training consumer's target mapping and inside the columns its loss reads. We read the released checkpoint itself on the forgiven scenes and report what it still recovers of the pre-overwrite distinction, beside a base-rate-matched control and the drivable-area head, whose column the rule never fires on (Section 6); Appendix I gives the prior work behind the head comparison. Continuing that checkpoint with matched seeds, we observe higher lane keeping on forgiven navtest scenes and lower adjacent-frame plan consistency under pre-overwrite than published targets, while the extended PDM score (EPDMS) with forgiveness disabled does not separate the supervision policies (Section 8).

## 2 Related work

**The rule and its evaluation-side audit.** Cao et al. (2025, Section 3.1) introduce and call novel the EPDMS filtering mechanism: a rule violation is forgiven when the human expert also commits it in the same scene. Wei et al. (2026, Sections 3.3 and 4.2) subsequently formalize reference-conditioned forgiveness and measure its evaluation-side interaction with numerical rollout failures. Their published audit protocol recommends the failure-mask and overwrite-rate measurements used here, and they measure evaluation-side overwrites for DAC, DDC, LK, TTC, and NC on 12,146 navtest tokens (Wei et al., 2026, Sections 4.2 and 5). The per-channel overwrite rates below are that recommended measurement, carried out on a different artifact.

Their unit is one scored trajectory per agent, so a channel holds a single value; the label generator writes 16,384 candidate values per column, and their rate is the one-entry case of the per-column magnitude measured here. The recommendation's stated purpose is to precede the interpretation of a ranking, whereas the object here is a released training-label file; their rates are measured under the rollout divergence they audit, and, with their thresholds read as deviation or step limits, no stock-solver human-reference rollout of this split exceeds them (Section 4 and Appendix L).

**Label-error audits.** Confident learning flags likely label errors from out-of-sample predicted probabilities (Northcutt et al., 2021b, Sections 2–3), and Northcutt et al. (2021a, Sections 3–4) validate such flags in ten benchmarks with crowdworkers; there the unit is one example judged against a latent true class, here a column of candidate values against its regenerated pre-overwrite side.

**Constant groups and ranking supervision.** The phenomenon class is established in both reinforcement learning with verifiable rewards and learning-to-rank. In GRPO, identical rewards yield zero group-relative advantages, as described by Coelho et al. (2026, Section 3.4, "Standard GRPO" and "Dynamic Pool Management"). XGBoost skips equal-label pairs and defines effective pairs by their contribution to the ranking gradient (XGBoost Contributors, 2025a, "Constructing Pairs" and "Obtaining Good Result"). Its default nDCG convention assigns 1 to a query with no positive labels, illustrating why a passing aggregate value need not attest to a ranking distinction (XGBoost Contributors, 2025b, `eval_metric`, `ndcg-`). Within driving, FlowR2A replaces nearly constant v1 comfort labels with history comfort to obtain more ranking signal (Li et al., 2026, Appendix F.2.4). Unlike the group-relative reward term, GTRS's component BCE can still have prediction-dependent gradients when its targets are constant.

**Abstention and positive supervision.** If forgiveness expresses that a rule cannot judge a scene, abstaining from that rule and assigning a positive target express different supervision decisions. Treating abstention as a first-class output of a labeling function is prior weak-supervision tooling (Ratner et al., 2017, Section 2).

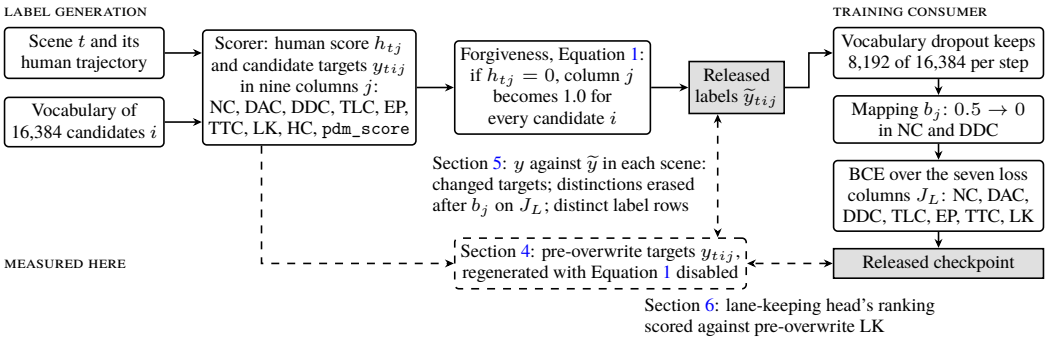


Figure 1: Label generation and the training consumer as released (solid), and the pre-overwrite side regenerated here with the comparisons made on it (dashed). Shaded boxes are the released label file and checkpoint.

**Whether an item still carries a target distinction.** Zimmerlin et al. (2024, Section 3, "Data filtering") prune an end-to-end driving dataset by a heuristic that estimates whether a frame changes the model's target labels relative to the previous frame. The predicate used here asks the same question of a different object: whether a forgiven column still separated candidates within one scene's vocabulary before it was overwritten, not whether a frame changes the targets relative to the previous frame. Both test target variation to decide whether an item still carries supervision, there to drop samples and here to read what an overwrite removed; the constant-group logic itself is the learning-to-rank one named above. The same work reports that adjusting how its CARLA expert brakes for pedestrians cut pedestrian collisions about fourfold for policies trained on the adjusted demonstrations, with the expert's own pedestrian collision rate unchanged (Zimmerlin et al., 2024, Section 3, "Expert style"). There the reference supplies the action being imitated, whereas here it gates the metric that writes the targets; what carries across the two settings is that a reference-side choice its own score does not register can reach what a policy is trained on.

**What this measurement adds.** What none of them reports is how this rule behaves on this benchmark's released training labels: how often it fires, and how much of a candidate vocabulary it reverses when it does.

## 3 Label generation and the training consumer

The audited GTRS-Dense file, `navtrain_16384.pkl`, has nine columns and 16,384 candidate entries per column: no-at-fault collisions (NC), drivable-area compliance (DAC), driving-direction compliance (DDC), traffic-light compliance (TLC), ego progress (EP), time to collision (TTC), lane keeping (LK), history comfort (HC), and the aggregate `pdm_score`. LK fails a candidate that stays more than 0.5 m from the route centerline for 20 consecutive steps, skipping steps inside intersections (Appendix F). Figure 1 follows one scene from the scorer that writes these columns to the loss that reads them. The source revisions are NAVSIM `0a380a9` and GTRS `92a740de`; we bind NAVSIM `v2.0`'s DDC body in place of GTRS's, whose intersection exclusion can only raise DDC, as the released labels follow the earlier body (NAVSIM Contributors, 2025; GTRS Contributors, 2026).

**The overwrite.** The `human_penalty_filter` branch in `pdm_score_full_v2` executes `gt[column] = np.ones_like(gt[column])` at `navsim/evaluate/pdm_score.py:308` when the separate human-scoring call returns zero for that column (GTRS Contributors, 2026). The assignment follows conversion to float16 and removal of the internal reference row from the candidate arrays. Writing $y_{tij}$ for the stored-precision pre-overwrite target, $h_{tj}$ for the human score, and $\widetilde{y}_{tij}$ for the published target, the rule is

$$\widetilde{y}_{tij} = \begin{cases} 1.0, & h_{tj} = 0, \\ y_{tij}, & h_{tj} \neq 0. \end{cases} \tag{1}$$

The branch is enabled on every scene of the measured split. The human reference used by this rule is separate from the internal reference used to normalize vocabulary progress. The upstream Issue #151

fix excludes `pdm_score` from human filtering and rebuilds aggregate inputs in the evaluation entry `pdm_score_from_interpolated_trajectory` (VVeiCao, 2025). The inspected GTRS label entry `pdm_score_full_v2` retains Equation 1, so that patch leaves the seven-column target overwrites measured here in place.

**Loaded fields and loss terms.** The loader takes the eight keys of `trajectory_pdm_weight` at `navsim/agents/gtrs_dense/gtrs_agent.py`:186 of `NVlabs/GTRS` commit `92a740de` and loads those fields at line 254; `pdm_score` is never loaded. HC is loaded but omitted from the loss sum at lines 129–138, leaving seven loss columns, $J_L = \{\text{NC}, \text{DAC}, \text{TTC}, \text{EP}, \text{DDC}, \text{LK}, \text{TLC}\}$. The mapping $b_j$ sends 0.5 to 0.0 for NC and DDC and is the identity for the other columns. A raw mixed column can therefore become constant before the loss reads it. The loss adds component-wise BCE terms rather than multiplying the component targets (GTRS Contributors, 2026, `hydra_kd_imi_agent_loss_dropout`).

**The earlier scene filter.** NAVSIM documents screening its navtrain list offline by a human-PDMS threshold of 0.8 under a v1 aggregate that contains neither LK nor TLC (Dauner et al., 2024, Section 3.1, "Filtering for challenging scenes"); Appendix B gives the filter's definition and tests it.

## 4 Measurement method

**Population and human census.** The denominator is the benchmark's own `navsim/planning/script/config/common/train_test_split/scene_filter/navtrain.yaml` token list, exactly 103,288 scenes. We use its complete original-scene population and the corresponding released GTRS-Dense labels. The human-zero mask is the label entry's own human call, scored without history. Human EP is identically 1.0 under singleton progress normalization, whereas human HC is identically 1.0 because this call omits history, so both human-zero sets are structurally empty. DAC's empty human-zero set is empirical: the v1 human census contains zero DAC-zero scenes but 20 NC-zero scenes in the published navtrain list. Both are multiplicative penalties, so enforcing the stated 0.8 criterion would remove both zero sets. The surviving NC zeros separate DAC's observed emptiness from that filter prediction.

**Solver sensitivity of the human census.** The census uses `np.linalg.pinv`, the stock pseudoinverse, for which Wei et al. (2026, solver matrix in Section 4.2) report a rollout-divergence rate of 99.8% on their own 450-scene interactive pool. Under the software configuration they audit, substituting `scipy.linalg.pinvh` or direct solve leaves all 103,288 tokens' nine human values bitwise identical. Under the deviation-or-step reading, 0 of the 103,288 stock-solver human rollouts of this split exceed the $(20, 10)$, $(100, 50)$, and $(200, 100)$ m thresholds; under the path-length reading, at least 5,165 exceed the strict 20 m limit and none exceeds 100 m; forgiveness erases a distinction in a loss column on 10.8793% of the same scenes (Appendix L). Neither these distances nor the solver table estimates a prevalence difference between the two corpora, whose channel sets, pools, and measured quantities all differ.

**Separate the predicates.** Let $J$ denote the nine stored columns and define $P_1(t)$ by the existence of $j \in J$ with $h_{tj} = 0$. Define $P_2(t)$ by a forgiven column containing at least one pre-overwrite value different from 1.0, and $P_3(t; K, b)$ by a forgiven column $j \in K$ containing at least two distinct values after mapping $b_j$. The raw definition uses the identity mapping; the loss-effective definition uses the consumer mapping from Section 3. Restricting to the loss columns removes 237 tokens: $J \setminus J_L$ holds only HC and `pdm_score` and the rule never fires on HC, so those 237 are exactly the scenes whose only loss-effective mixed forgiven column is the aggregate. The main count is the token union $P_3(t; J_L, b)$. For each forgiven column we record whether it was already all ones, constant at a failing value, or mixed before overwrite. A column can lose distinctions without the scene's complete label matrix becoming constant.

**Pre-overwrite reconstruction and paired rows.** We regenerate the candidate arrays with the final forgiveness assignment disabled, retain the released float16 precision, and join the arrays to the published labels by token and candidate index. The assignment is the generator's last step, so these values come from the computation that wrote the released file. The pre-overwrite values appear in no

Table 1: Nested token predicates over the official 103,288-scene navtrain split. $P_1$ and $P_2$ share a row because no forgiven column is already all-passing before the overwrite.

| Predicate | Tokens | Fraction (%) |
|---|---|---|
| $P_1 = P_2$: human-zero trigger, and the overwrite changes a value | 12,576 | 12.1757 |
| $P_3$: raw mixed, any stored column | 11,503 | 11.1368 |
| $P_3$: loss-effective mixed, any stored column | 11,474 | 11.1087 |
| **$P_3$: loss-effective mixed, a loss column** | **11,237** | **10.8793** |

released artifact, so we check the reconstruction several ways. Within the $P_1$ pool, every column the overwrite did not touch is compared against the published file, and 97,173 of 97,175 comparisons are equal; both exceptions are the non-forgiven EP and aggregate columns of one token. That denominator is $12{,}576 \times 9$ less the 16,009 forgiven token-columns. A 60-token control pool on which the rule cannot fire is bitwise equal on all nine columns, 540 of 540 token-column comparisons. The scorer's own per-metric output, captured in a separate run with forgiveness on, gives the same DDC value-set partition and the same DDC counts at the five ranks as the reconstruction (Appendix J). A corpus-wide census of the two arrays finds a difference on all 12,576 tokens, and the nine tokens that differ without belonging to the $P_1$ pool differ only on ego progress, a column forgiveness never fires on, and the aggregate it feeds. Run with forgiveness on, the same configuration reproduces all nine published columns byte for byte on 103,278 of 103,288 tokens (Appendix C), which also localizes the cache dependency behind 8 of the 10 residuals. We count distinct rows in three column representations: all nine raw columns, the seven raw loss columns, and those seven columns after NC/DDC binarization.

## 5 Forgiveness in the released training labels

### 5.1 Every trigger changes a target, and most also erase a distinction

Table 1 separates the human-zero trigger from its two possible label consequences. To scale its last row we measured a second rewrite these labels undergo on the same 103,288 scenes: before the BCE the loss maps 0.5 to 0 in the collision and driving-direction columns, and on 260 scenes, 0.2517%, that mapping leaves a loss column constant which the published values had left mixed. On this split, forgiveness erases distinctions on 43.2× as many scenes as the NC/DDC target mapping. The two overlap on 73 scenes.

**The trigger union is preserved structurally.** The 2,967 aggregate-column forgivenesses are not read by this loss, but cannot be the sole trigger: human EP and HC are both 1.0, making the human weighted factor positive, so a zero aggregate requires a zero in NC, DAC, DDC, or TLC, all of which enter the loss. HC itself cannot trigger forgiveness, so restricting the trigger to the seven loss columns necessarily retains all 12,576 members of $P_1$.

### 5.2 Lane keeping accounts for most forgiven loss columns, and firing usually reverses most of a column

LK forgiveness accounts for 79.37% of the $P_1$ pool. On 9,831 navtrain scenes (9.52%), LK is overwritten to passing for all 16,384 candidates despite distinguishing passing and violating candidates before overwrite. LK's observed value sets are exactly $\{0.0, 1.0\}$ for the mixed and $\{0.0\}$ for the all-zero columns of Table 2; LK bypasses binarization, so its raw and loss-effective classifications coincide. Among mixed LK columns ($n = 9{,}831$), violating candidates have minimum 306, median 14,323 of 16,384 (87.42%), and maximum 16,383; including the 151 all-zero columns gives a forgiven-pool median of 14,391 ($n = 9{,}982$). Figure 2 shows the whole distribution. A candidate that fails lane keeping keeps its failing target on the 86,107 unforgiven scenes and receives a passing one on the 9,831 forgiven ones; in the scorer's own rollouts the entries turned to passing reach a median of 6.09 m from the route centerline at their farthest, against 7.96 m for those kept failing, and their longest run beyond the lane-keeping limit has a median of 33 against 26 of 41 steps (Appendix F).

Table 2: Pre-overwrite columns after the consumer mapping. "Mixed" means at least two distinct target values, and the last column pools every forgiven token-column of that column, all-zero and mixed together, and gives the upper median number of candidates that failed it before the overwrite (Appendix J). The aggregate is a diagnostic field this consumer's loss does not load.

| Column | Forgiven | Already one | All zero | Mixed | Mixed (% of forgiven) | Pooled median failing |
|---|---|---|---|---|---|---|
| LK | 9,982 | 0 | 151 | 9,831 | 98.49 | 14,391 |
| TLC | 2,156 | 0 | 1,057 | 1,099 | 50.97 | 16,380 |
| DDC | 876 | 0 | 386 | 490 | 55.94 | 16,317 |
| TTC | 18 | 0 | 9 | 9 | 50.00 | 16,384 |
| NC | 10 | 0 | 5 | 5 | 50.00 | 16,384 |
| DAC (empirical empty set) | 0 | 0 | 0 | 0 | — | — |
| EP (structural empty set) | 0 | 0 | 0 | 0 | — | — |
| HC (structural empty set) | 0 | 0 | 0 | 0 | — | — |
| *Seven loss columns* | *13,042* | *0* | *1,608* | *11,434* | | |
| `pdm_score` (unloaded) | 2,967 | 0 | 1,179 | 1,788 | 60.26 | 16,384 |

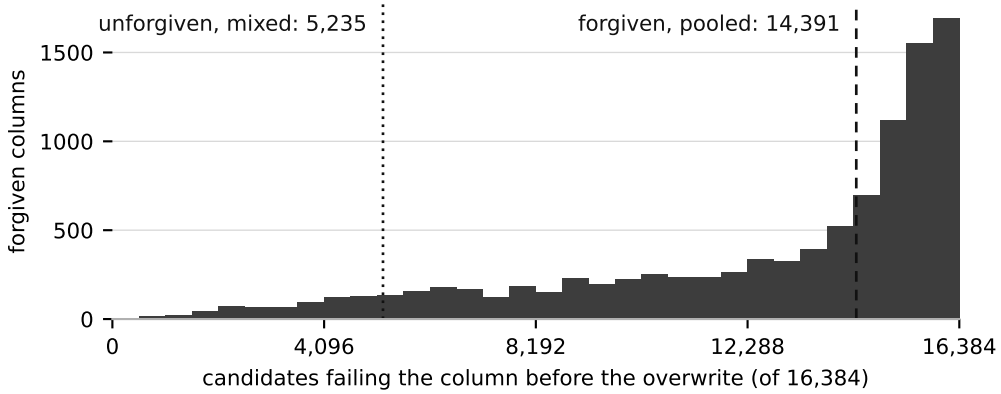


Figure 2: Forgiven lane-keeping columns by how many of the 16,384 candidates failed the column before the overwrite, one count per column ($n = 9{,}982$, bin width 512). The mass sits against the upper end: where the rule fires on this column, it usually replaces one in which most candidates were failing. The dotted line is the same count where it does not fire, median 5,235 over the 86,107 scenes with a mixed published column; the corresponding mixed-column median on the forgiven side is 14,323, since the histogram also carries the 151 columns that were already constant. Lane keeping lies below each of the other four firing loss columns at the lower four quantiles (Appendix J).

**Scale in supervision entries.** Changed LK targets total 126,696,155 entries, 124,222,171 in mixed and 2,473,984 in all-zero columns, out of $103{,}288 \times 16{,}384 \times 7 = 11{,}845{,}894{,}144$ target entries. The changed entries in mixed LK columns alone are 1.0487% of all target entries. The other four triggered loss columns add 43,353,063, so the rule changes 170,049,218 target entries in the seven loss columns, 1.4355% of them (Appendix J). The released BCE weights for NC, DAC, TTC, EP, DDC, LK and TLC are $(3, 3, 4, 2, 1, 2, 3)$, so LK has $2/18$ of their sum; a separate imitation term enters with `trajectory_imi_weight` (default 1.0). These entry counts and loss weights do not determine gradient shares (Section 2).

**Raw and loss-effective readings.** Only NC and DDC change under the half-to-zero mapping: mixed NC columns stay at five, while mixed DDC columns fall from 863 of 876 raw (98.52%) to 490 of 876 loss-effective (55.94%). Forgiveness also fires on the v1 channels NC and TTC, on 10 and 18 tokens respectively, so its observed support is not confined to later-added channels.

Under the released vocabulary dropout, both pre-overwrite target values survive with mean probability 0.998323, weighting each of the 11,434 mixed forgiven loss token-columns equally (Appendix E).

Table 3: Paired distinct-row counts over the $P_1$ pool of 12,576 tokens. Medians are computed separately before and after overwrite; the mean drop averages per-token differences. Dropping the one token whose untouched columns differ (Section 4) lowers each strict-decrease count by one and moves no entry by more than 0.0089%.

| Representation | Pre median | Post median | Mean drop | Strict decreases |
|---|---|---|---|---|
| All nine raw columns | 1,318 | 1,293 | 20.50 | 11,499 (91.44%) |
| Seven raw loss columns | 1,214 | 1,148 | 121.54 | 11,502 (91.46%) |
| Seven loss-effective columns (binarized) | 1,205 | 1,142 | 118.99 | 11,236 (89.34%) |

The repository's Hydra-MDP configuration takes the same file as its training labels with vocabulary dropout off, so every candidate enters its loss (GTRS Contributors, 2026, `hydra_mdp_vov.yaml`).

### 5.3 Overwrite merges label rows on most tokens

Table 3 compares each token with itself, using the reconstructed pre-overwrite and published arrays defined in Section 4; pairing within a token is what keeps selection into forgiveness from producing the effect, since both sides of every comparison are the same scene. The loss-column projection exposes more row mergers than the full stored representation, whose extra coordinates can keep apart rows the projection merges. The 90,712 untouched tokens have median 1,627 and mean 1,692.52 distinct nine-column rows; the forgiven pool already has approximately 300 fewer at the median before overwrite (Table 3). This between-pool contrast describes selection into forgiveness.

In all three representations, no token has a single label row before or after overwrite. After loss binarization, the 11,236 tokens with fewer distinct rows all lie in the 11,237 of Table 1's last row; in the remaining token a forgiven loss column is mixed, but its distinction is fully redundant with the other six columns, so the number of distinct rows is unchanged.

## 6 What the released weights recover where the rule fired

We also read the released GTRS-Dense checkpoint `gtrs_dense_vov.ckpt`, the artifact the publisher's training configuration pairs with this label file, ranking every scene's 16,384 candidates by the lane-keeping head's logit and scoring that ranking against the pre-overwrite lane-keeping outcome. Medians over the 9,831 mixed LK-forgiven scenes, 95% intervals bracketed and rounded outward: AUC 0.6538 [0.6465, 0.6609], average precision 0.201, McClish partial AUC below a 0.1 false-positive rate 0.516. Its fifth and twenty-fifth percentiles both sit exactly at the McClish floor 0.474, which a partial AUC of zero maps to, so at least a quarter of those scenes have a zero partial AUC. Most of what the forgiven group recovers is matched by a head without lane-keeping supervision: the same pass's drivable-area logits, read against the same labels, reach $\mathrm{AUC}_{\mathrm{DAC}\to\mathrm{LK}} = 0.6241$ where the head reaches 0.6538 on the forgiven scenes, and 0.8489 where it reaches 0.9940 over all 86,107 unforgiven scenes with a mixed published column, covering 81% and 71% of its excess over chance. The overwrite changed no drivable-area value on this split, so the shares compare the lane-keeping head with a head of the same network whose own targets it never touched, on the same scenes and base rates. Matching one-to-one on the failing-candidate count pairs 7,940 of those scenes with an unforgiven scene and leaves 1,891 unmatched. For $D$ the within-scene $\mathrm{AUC}_{\mathrm{DAC}\to\mathrm{DAC}} - \mathrm{AUC}_{\mathrm{LK}\to\mathrm{LK}}$, each head scored against its own column, the primary contrast $D_{\text{forgiven}} - D_{\text{matched}}$, whose drivable-area term is the control for how hard a scene is to rank overall, is $+0.2850$ [0.2770, 0.2922] over those pairs and $+0.3710$ [0.3658, 0.3766] under direct standardization over strata of that count. As a secondary reading, over the pairs the unforgiven side reaches AUC 0.9905 [0.9901, 0.9908], average precision 0.956 and McClish partial AUC 0.953, against an AUC of 0.7046 on the 7,940 matched LK-forgiven scenes, a difference of medians of $-0.2859$ [$-0.2939$, $-0.2797$] (Appendix H). Discriminability falls on forgiven scenes but stays above chance. Labels differ in source between the groups: published labels on the matched one and our reconstruction (Section 4) on the forgiven one. The gap holds on held-out navtest scenes, whose forgiven and matched groups take their labels from one regeneration: the median AUC is 0.7095 [0.6847, 0.7241] on all 1,496 mixed LK-forgiven scenes against 0.9807 [0.9780, 0.9820] on the 1,287 unforgiven scenes matched to them, and the primary contrast is $+0.2368$ [0.2206, 0.2544] (Appendix K). The Hydra-MDP checkpoint released

with GTRS, whose configuration takes the same file (Section 5.2), shows the gap too: 0.6627 against 0.9761 on navtest, with primary contrasts of $+0.2571$ on navtest and $+0.3310$ on navtrain, both intervals above zero (Appendix H).

## 7 Label regeneration depends on the OpenBLAS kernel selected for the processor

Under the same code, metric cache and scoring configuration that reproduce the released GTRS-Dense labels bitwise on Broadwell across the full corpus except ten residuals in two of the nine columns, regeneration on Intel Sapphire Rapids matches all nine columns jointly on 0 of 4,992 tokens and on 0 of a separate 260-token sample, whose regenerated arrays a second machine matches bitwise. With the default kernel, the maximum absolute difference on those 4,992 tokens reaches 1.0 on every one of the nine columns, whose values lie in $[0, 1]$ and seven of which are restricted to $\{0, 1/2, 1\}$, so in every column at least one label flips rather than differing in the last bit; the share of a token's nine-column entries that differ from the label file has median 51.5% and maximum 77.7%. Schlögl et al. (2023) find that across 75 platforms the deviations in a trained model's softmax outputs rarely change a label on natural inputs; on the dual-core x86 systems available on Google Cloud, the CPUs fall into equivalence classes aligned with clusters of co-occurring SIMD flags, one holding `avx512f`. Here the object is the released label file, and a token matches only when its complete nine-column array does. The same symptom has been on the benchmark's issue tracker since 2024, attributed there to `np.linalg.pinv` and to an OpenBLAS kernel on newer Intel parts, with the maintainer's question of whether two reports share a cause unanswered; they compare scores, not artifacts, and do not separate the CPU from the accelerator, and the readings above are that comparison made on the label file (Appendix G). Three repeated Sapphire Rapids runs of the same 52 tokens agree bitwise with each other, and the nine columns are invariant across two Sapphire Rapids SKUs, matching the CPU determinism Schlögl et al. report; the difference is between platforms and not between runs. The difference enters through the OpenBLAS library bundled with NumPy, which selects its kernel from the processor at run time and selects a different one on Sapphire Rapids than on Broadwell; with that library alone pinned to the kernel it selects on Broadwell, regeneration on Sapphire Rapids matches all nine columns on 4,991 of the 4,992 tokens, and on all 4,992 with every OpenBLAS build pinned and NumPy's AVX-512 dispatch disabled (supplementary material). On the separate sample, AVX-512 alone does not produce the difference: with NumPy's and SciPy's builds on SkylakeX, also an AVX-512 kernel, regeneration on Sapphire Rapids matches all nine columns on all 260 tokens, where the default Cooperlake kernel matches none. Feeding Broadwell's simulator output into the Sapphire Rapids scorer recovers the released labels bitwise on all eight tokens tested. On the 4,992 tokens, which include those eight, the trajectory array entering the simulator is bitwise the same under the default and the pinned kernel while the labels differ on every token (supplementary material); together with the cross-feed on the eight, this places the divergence inside the simulator.

## 8 Pre-overwrite targets shift lane keeping and extended comfort, more on forgiven scenes

We continue the released checkpoint from step 4,800, including its Adam state, for 300 optimizer steps on navtrain with 3 seeds per supervision: A retains the published targets, M omits forgiven token-columns from their component loss, and P uses pre-overwrite targets. Evaluation uses single-stage non-reactive simulation on 12,146 navtest tokens: OFF disables forgiveness, and ON keeps it and rebuilds the aggregate after human-zero overwrites, as the upstream correction does (VVeiCao, 2025). The benchmark's extended comfort (EC) scores consistency between adjacent-frame plans (GTRS Contributors, 2026, `ego_is_two_frame_extended_comfort`), over tokens with an adjacent frame. F holds the 1,816 navtest tokens whose human reference scores zero on LK, DDC, TLC, NC or TTC, and R the other 10,330. A contrast with A at the same seed is observed (O) when all seeds share its sign, every 95% interval excludes zero, and its absolute mean exceeds the largest gap between seeds of one supervision, read over seeds 0–1 and 0–2 (the two- and three-seed readings).

We observe higher OFF lane keeping and lower EC for P than A, with both changes larger on forgiven scenes (F) than on the rest (R): the LK interaction is positive and the EC interaction negative in every seed, with intervals excluding zero (Table 4). OFF lane-keeping gains on F are +2.478/+3.524/+3.414

Table 4: P–A subset interactions in score points, with 95% intervals: the paired difference on F minus that on R. Both are observed (O) in the two- and three-seed readings (Appendix K).

| Metric | $s = 0$ | $s = 1$ | $s = 2$ |
|---|---|---|---|
| LK OFF | +1.820 [+0.619, +3.134] | +3.243 [+1.704, +4.861] | +2.795 [+1.377, +4.152] |
| EC | -5.740 [-8.112, -3.385] | -5.583 [-8.070, -3.109] | -4.236 [-6.772, -1.658] |

points, versus +0.658/+0.281/+0.620 on R, in seed order. On F, where the human reference itself fails a check, these gains come with a centimeter-scale shift from the human trajectory, 0.050–0.084 m on average (ADE); on R the distance does not separate. The EC decline also appears on the tokens whose lane-keeping outcome does not change, by 1.448–2.829 points with an interval below zero in every seed. M's all-token OFF lane-keeping contrast is also observed (O), while its subset contrasts do not separate. For the positive control L0, with every lane-keeping target passing, we observe lower OFF lane keeping than A at their shared seed: this training budget registers a lane-keeping response. On Section 6's probe, the in-sample median AUC on mixed LK-forgiven scenes is 0.9225–0.9233 for P against 0.6141–0.6423 for A, and the primary contrast 0.0563–0.0570 against 0.3007–0.3225; on held-out navtest that AUC is 0.1086–0.1251 higher for P than for A, and the primary contrast 0.1234–0.1399 against 0.2249–0.2341 (Appendix K). At 530 steps, P–A's OFF lane-keeping and EC contrasts keep their signs and verdicts on all tokens and on F, as do both interactions and the probe contrasts of Section 6, and OFF EPDMS still does not separate P from A (Appendix K). Lane keeping and comfort separate P from A, by more on forgiven scenes than on the rest, while OFF EPDMS, the aggregate that includes both, does not separate the supervision policies (Appendix K).

## 9 Discussion

Because EPDMS applies the same rule when it scores, the overwritten targets agree with the benchmark's own evaluation; what the overwrite removes, where a forgiven column was mixed, is the pre-overwrite ordering among a scene's candidates, which is the quantity measured here. Consistent with this alignment, we observe higher ON EPDMS under published than pre-overwrite targets at 300 steps in the two-seed reading only, and at 530 steps in neither reading (Appendix K). Removing forgiveness has its own cost: Wei et al. (2026, Section 5; Table 4, stock filter-off row) document zero human-replay and PDM-Closed planner scores on interaction-heavy candidates and near-zero corpus means in their stock filter-off control, whose rollouts diverge (Section 4). The all-excluded $B$ class in Appendix D supplies a label-side instance of this cost on navtrain. The overwrite is not inert for training: with the pre-overwrite targets restored, we observe higher lane keeping on forgiven navtest scenes, while OFF EPDMS does not separate the supervision policies (Section 8).

**Limitations.** The continuation covers GTRS-Dense on navtest, starting from weights trained with forgiveness. The v2 label census and solver contrasts run on an Intel Xeon E5-2640 v4 (Broadwell); the separate v1 human census spans the five CPU models recorded in Appendix A. The navtrain mixed-versus-constant classification counts were taken on Broadwell only, not on AMD Rome, where Appendix F's lane-keeping recomputation equals the stored pre-overwrite column on all 67,035 tokens run there, or on the Sapphire Rapids parts of Section 7, where the published labels do not reproduce at all under the default kernel. With forgiveness disabled there, none of 832 tokens matches the Broadwell reconstruction on all nine columns. The label census covers navtrain, not navtest, navhard, or synthetic frames. The exposure result (Appendix E) is a combinatorial expectation over the sampling distribution; no actual training run's dropout indices were read. The rollout-divergence readings cover the human-reference rollout under the stock solver only.

## AI use statement

In this work, we used generative AI tools to search for, identify and summarize relevant literature, and to edit and polish the text of this paper for readability. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, to interpret results, to propose or refine hypotheses, to design or give feedback on research methodology or experiments, or to implement methods. Generating synthetic data sets, formulating mathematical claims, providing critical ingre-

dients for proving mathematical claims, assisting in the writing of proofs, translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for the creation of artifacts. We have reviewed all AI-assisted work. The papers we cite were read in full, and a citation was checked at the place in the cited paper that supports the claim it carries; the authors edited the text and verified it against the results it reports. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## Reproducibility statement

The supplementary material gives the settings behind its numbers that the paper does not state: the public inputs with their digests, the configurations of the label-regeneration, metric-cache and checkpoint-reading runs, the token samples and processors of Section 7, the continued-training and evaluation configurations of Section 8, the order-statistic conventions, the rollout settings of Appendix F, and the resampling, tie treatments and held-out labels behind the readings of Section 6. Appendix C reports the full-corpus certification of the published labels and the AMD Rome sample, and Section 7 the Sapphire Rapids differences.

## References

Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. Data validation for machine learning. In *Proceedings of Machine Learning and Systems*, volume 1, pp. 334–347, 2019.

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, 2025. URL https://arxiv.org/abs/2506.04218v3. Version 3, 20 March 2026.

João Coelho, João Magalhães, Bruno Martins, and Chenyan Xiong. Effective reinforcement learning for agentic search by recycling zero-variance queries during training, 2026. URL https://arxiv.org/abs/2606.10709v1. Version 1, 9 June 2026.

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, 2024. URL https://arxiv.org/abs/2406.15349v2. Version 2, 31 October 2024.

GTRS Contributors. GTRS source code. GitHub repository, NVlabs/GTRS, 2026. URL https://github.com/NVlabs/GTRS/tree/92a740def80610e4096962d25cc837b21e72ef78.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://arxiv.org/abs/1909.03368v1. Version 1, 8 September 2019.

Xirui Li, Zhe Liu, Xiaoqing Ye, Wenhua Han, Yifeng Pan, Junyu Han, and Hengshuang Zhao. FlowR2A: Learning reward-to-action distribution for multimodal driving planning, 2026. URL https://arxiv.org/abs/2606.24231v1. Version 1.

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, 2024. URL https://arxiv.org/abs/2406.06978v4. Version 4, 30 August 2024.

Zhenxin Li, Wenhao Yao, Zi Wang, Xinglong Sun, Joshua Chen, Nadine Chang, Maying Shen, Zuxuan Wu, Shiyi Lan, and Jose M. Alvarez. Generalized trajectory scoring for end-to-end multimodal planning, 2025. URL https://arxiv.org/abs/2506.06664v1. Version 1.

Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M. Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving?, 2023. URL https://arxiv.org/abs/2312.03031v2. Version 2, 2 June 2024.

Christian Löwens, Thorben Funke, and Alexandru Paul Condurache. Driving on memory, 2026. URL https://arxiv.org/abs/2608.31029v1. Version 1, 31 August 2026.

NAVSIM Contributors. NAVSIM source code. GitHub repository, autonomousvision/navsim, 2025. URL https://github.com/autonomousvision/navsim/tree/0a380a9063d7162ec93d0f51e9990ebac585f720.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021a. URL https://arxiv.org/abs/2103.14749v4. Version 4, 7 November 2021. NeurIPS 2021 Track on Datasets and Benchmarks.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021b. doi: 10.1613/jair.1.12125. URL https://arxiv.org/abs/1911.00068v6.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision, 2017. URL https://arxiv.org/abs/1711.10160v1. Version 1. Proceedings of the VLDB Endowment 11(3), 269–282, 2017.

Alexander Schlögl, Nora Hofer, and Rainer Böhme. Causes and effects of unanticipated numerical deviations in neural network inference frameworks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56095–56107. Curran Associates, Inc., 2023. doi: 10.52202/075280-2446.

VVeiCao. Fix human filter bug (Issue #151). Commit to autonomousvision/navsim, September 2025. URL https://github.com/autonomousvision/navsim/commit/359c7f72304bfa8273e754224a213d3751bd2340. Commit 359c7f7; changelog date: September 29, 2025.

Ziang Wei, Minjun Yu, Zheyuan Lai, Mingjie Pang, and Wei Li. When shared rollouts fail in defensive driving evaluation: A NAVSIM score basis audit, 2026. URL https://arxiv.org/abs/2608.04896v1. Version 1.

XGBoost Contributors. Learning to rank. XGBoost documentation, release 3.0.0 documentation branch, 2025a. URL https://xgboost.readthedocs.io/en/release_3.0.0/tutorials/learning_to_rank.html. Sections: Constructing Pairs; Obtaining Good Result.

XGBoost Contributors. XGBoost parameters. XGBoost documentation, release 3.0.0 documentation branch, 2025b. URL https://xgboost.readthedocs.io/en/release_3.0.0/parameter.html. Learning Task Parameters: eval_metric, ndcg-.

Julian Zimmerlin, Jens Beißwenger, Bernhard Jaeger, Andreas Geiger, and Kashyap Chitta. Hidden biases of end-to-end driving datasets, 2024. URL https://arxiv.org/abs/2412.09602v2. Version 2.

## A Reconstructed v1 human scores

The census scores all 103,288 tokens, and the eight consistency checks listed in the supplementary material hold on every token. It spans Intel E5-2640 v4 and Gold 5118, and AMD EPYC 7502, 7543, and 8224P. The paired means, EP 0.8799 and PDMS 0.9495, sit near the published navtest human row's 0.875 and 0.948 (Dauner et al., 2024, Table 1, human row). This anchor supports the reconstruction, not the filtering. The cache was built by v2, whose observation-construction chain differs from v1's. Its missing forecast second requires a TTC clamp. The measured TTC is consequently an upper bound on the intended v1 value. The v2 census also removes agents overlapping the ego at frame zero, whereas v1 uses non-reactive replay without that removal. On the 12 of 103,288 tokens where the v1 and v2 NC values differ, v1 is the lower; that input difference, rather than changed NC semantics, accounts for this direction.

**The v1 driving-direction member (Appendix B) is inert on this corpus.** The v1 loop slices the proposal axis rather than summing the intended one-second temporal window. DDC equals 1.0 on all 103,288 human trajectories, including all 2,003 that v2 scores below 1.0. In the artificial batch [ref, ref], the second row's accumulated vector is exactly twice the first row's on all 876 tested tokens, with 513 nonzero cases; the doubling crosses a scoring threshold on 2 of 876. The human DDC score changes between the single-row and paired batches on zero tokens. Thus v1's driving-direction member cannot fire on this human corpus, while the mechanism that would make it batch-dependent is measured in the artificial configuration.

## B Counterexamples to the documented split criterion

**The documented filter.** NAVSIM's split-generation paragraph names navtrain and navtest, removes scenes with human PDMS below 0.8, also removes scenes with constant-velocity PDMS above 0.8, and reserves adjustment of these thresholds to obtain the desired dataset size (Dauner et al., 2024, Section 3.1, "Filtering for challenging scenes"). Its published v1 equation uses $\{\mathrm{NC}, \mathrm{DAC}, \mathrm{EP}, \mathrm{TTC}, \mathrm{C}\}$ and omits DDC, whereas the released v1 code includes DDC as a multiplicative member; LK and TLC are absent from both v1 aggregates. A cited description of v1 records DDC instead as a weighted term of weight zero (Li et al., 2026, Appendix E.3); the two readings agree on this corpus because DDC is 1.0 on every human trajectory in it (Appendix A). The distributed `SceneFilter` has twelve fields, none expressing a score, metric, or threshold; its token-list consumption path checks set membership (NAVSIM Contributors, 2025, `SceneFilter`; `filter_scenes`). The documented score screening is an offline step embodied in the supplied list, outside the distributed split-consumption code, which cannot reapply it. No split-list generation code was found in the `autonomousvision/navsim` repository at `0a380a9` (supplementary material). New fields falling outside an older validation predicate are a familiar data-validation problem (Breck et al., 2019). The relationship between NAVSIM's original split filter and its later metric channels is an instance of this problem.

**Counterexamples.** Scoring all 103,288 navtrain human trajectories with the reconstructed v1 pipeline yields 34 scenes (0.0329%) below 0.8 even under the most permissive of the four aggregations, NAVSIM's published formula scored on the human trajectory alone. These counterexamples hold under both the published and the code aggregation and any choice of reference row. With the cached PDM-Closed trajectory as the reference row, the code aggregation gives 364 (0.3524%). Both counts are inconsistent with the described removal of every scene below 0.8; neither establishes what threshold, if any, was executed. The additional 330 are EP-only cases; their reference trajectory is the cached v2 PDM-Closed rollout, and the median factor needed to raise normalized EP to the threshold is 1.102. A 10% reduction of the denominator can move that median case across the threshold; the 34 reference-independent cases carry the conclusion without this dependence. Across the paired counterexamples, the nonexclusive failing-member counts are TTC zero 33, NC zero 20, NC half-valued 3, DAC zero 0, and DDC below one 0. The counts are lower bounds under three one-sided relaxations: missing TTC anchors in the v2 cache, truncation from v1's anchor count to v2's, and omission of the DDC factor in the published-formula one-row reading. The nearest measured score is $3.21 \times 10^{-6}$ from 0.8, with none within $10^{-6}$. NC's at-fault classifier is AST-identical between versions, and its score function differs only cosmetically. The changed definitions are DDC's aggregation axis and TTC's anchor count (41 versus 32). The `COMFORTABLE` member of NAVSIM v1.0 (`9f3ea68`) and v2's TLC, LK, HC, and `two_frame_extended_comfort` are not mapped between versions. Appendix A records the reconstruction controls and input boundaries.

## C Full-corpus certification and the cache behind eight of its residuals

**Full-corpus certification.** With forgiveness enabled, the full-corpus certification reproduces all nine published columns byte for byte on 103,278 of 103,288 tokens. The 10 mismatches split into 2 storage-precision ties and 8 cache-content residuals. This appendix localizes the cache dependency behind the latter. A storage-precision tie is a token whose `pdm_score` column is bitwise equal and whose every differing entry differs by one float16 unit in the last place at its value. That certification ran on Broadwell, and two further microarchitectures were tested on a separate 260-token sample: AMD Rome reproduces all nine columns on 260/260 tokens, and Sapphire Rapids is the subject

of Section 7. Pinning every run of a comparison to one processor keeps the processor out of the differences between those runs, which is a separate matter from whether the labels reproduce on it.

The eight cache-content residuals trace to accumulated red-light identifiers consumed during leading-agent selection for the Intelligent Driver Model (IDM) in `autonomousvision/navsim@0a380a9`, `navsim/planning/simulation/planner/pdm_planner/proposal/pdm_generator.py`:263. Clearing the accumulated field or short-circuiting that red-light use recovers seven of the eight; the remaining residual, `9430eba18cb05d7c`, is reproduced by injecting its measured identifier set, and a log-grouped construction, one chunk per log, reproduces all eight. The log-grouped construction identifies a configuration because it changes both chunk size and composition; the field and consumption-site interventions localize the mechanism.

## D Label matrices and candidate-set overlap

Before applying any selection statistic, the two label matrices are bitwise identical on all nine columns for 90,703 of 103,278 tokens after excluding the ten reproduction residuals. The affected pool here is exactly the complement, 12,575 tokens: the $P_1$ pool less its one reproduction residual. Over the full 103,288-token comparison, differing-column counts are LK 9,982, `pdm_score` 2,975, TLC 2,156, DDC 876, TTC 18, NC 10, DAC 0, and HC 0; EP differs on the ten reproduction residuals. The two accounts of `pdm_score` add rather than overlap: the 2,967 forgiven columns and the eight residuals whose `pdm_score` differs are disjoint sets, and their union is exactly the 2,975 columns on which the two matrices differ. The column reproduces bitwise on 103,280 of 103,288 tokens, failing on 8 of the ten and on neither storage-precision tie.

**Selection rule and set overlap.** The source consumes labels as seven per-column BCE targets and contains no argmax over the label matrix. Its vocabulary-wide argmax acts on student predictions at inference. We substitute the loss-mapped labels $t_j$ into that rule's metric part, a diagnostic construction rather than an operation performed by the training code:

$$S(i) = 0.1 \ln t_{\mathrm{TLC},i} + 0.1 \ln t_{\mathrm{NC},i} + 0.9 \ln t_{\mathrm{DAC},i} + 0.2 \ln t_{\mathrm{DDC},i} + 6 \ln(7t_{\mathrm{TTC},i} + 7t_{\mathrm{EP},i} + 3t_{\mathrm{LK},i}). \tag{2}$$

We promote stored float16 labels to float64, map NC/DDC as the loss does, retain $\ln 0 = -\infty$, and compare exact maximizer sets $M_A, M_B$ for published ($A$) and pre-overwrite ($B$) files without breaking ties. Evaluated on the label matrices with the metric part of the rule the released agent applies at inference (`navsim/agents/gtrs_dense/hydra_model.py`:264–271), its imitation term absent because no label column corresponds to it, the published and the pre-overwrite label files place different candidate sets first on 10,193 of 103,278 navtrain tokens (9.87%), and share no top-ranked candidate on 967 (0.94%). The set-different total is 81.06% of the affected pool; the decomposition follows with both denominators:

| Relationship | Tokens | % of 12,575 | % of 103,278 |
|---|---|---|---|
| $M_B \subsetneq M_A$ | 7,637 | 60.73 | 7.39 |
| $M_A \subsetneq M_B$ | 1,589 | 12.64 | 1.54 |
| Crossing | 0 | 0.00 | 0.00 |
| Disjoint | 967 | 7.69 | 0.94 |

The first class leaves a smaller jointly optimal set under $B$, refining the published tie without implying a unique candidate. In the second, $B$ excludes all 16,384 candidates: its formal maximizer set is the whole vocabulary at $-\infty$, and it ranks nothing. That class is 1,589 of the 10,193, so the set-different total excluding it is 8,604 tokens, 68.42% of the affected pool.

**Ties and the separate unique-candidate row.** The analysis is tie-dominated: both sets are singletons on 542 of 12,575 affected tokens (4.31%), both are nonsingletons on 86.85%, and their median sizes are 660 ($A$) and 130 ($B$). Tokens on which both sets are singletons and differ:

| Case | Tokens | % of 542 | % of 12,575 | % of 103,278 |
|---|---|---|---|---|
| Both unique, different | 167 | 30.81 | 1.33 | 0.1617 |

For these tokens, scoring both candidates with $A$'s aggregate gives median $\Delta^{(A)} = +11.91$ points, positive on 164/167; scoring both with $B$'s gives median $\Delta^{(B)} = +11.91$, positive on 165/167. The two medians are equal before rounding. Each difference is $B$-ranked minus $A$-ranked, on the 0–100 label scale, per scene and candidate.

**Other selection statistics.** Under the loss's own weights $w_j$ (Section 5.2), $S_{\text{lossw}} = \sum_j w_j t_j$, the split by whether the two files' maximizer sets differ agrees with that of $S$ on 10,971 of the 12,575 affected tokens (87.24%), and under $S_{\text{pdm}} =$ `pdm_score` on 3,180 (25.29%); the set-overlap counts above are those of $S$.

## E The erased distinction would usually survive vocabulary dropout

**Expected exposure under vocabulary dropout.** The released training configuration uses `randperm` and `num_total//2` to retain $s = 8{,}192$ of $N = 16{,}384$ candidates without replacement; the labels are sliced by the same indices. For a mixed binary column with $k$ candidates in one class, the standard hypergeometric calculation gives

$$p(k) = 1 - \frac{\binom{k}{s} + \binom{N-k}{s}}{\binom{N}{s}}, \qquad \binom{a}{s} = 0 \text{ for } a < s. \tag{3}$$

This is the probability that a sampled sub-vocabulary would contain both pre-overwrite target values. We pool it over mixed forgiven loss columns, weighting each token-column equally; a column already constant before overwrite has no distinction for dropout to remove. The constant published column remains constant under the same sampling. Across those 11,434 mixed forgiven token-columns in the loss, the mean probability in Equation 3 is 0.998323. For a mixed binary column, retaining half the candidates gives $p(k) \geq s/N = 0.5$, with equality at a singleton minority; probabilities below 0.5 are structurally excluded. For 11,084 of these 11,434 columns, the probability equals 1.0 to double precision. Averaged over these columns, the erased distinction would therefore be present on essentially every sampled appearance of the corresponding scene during training. The calculation closes one alternative: if dropout usually removed the distinction anyway, the overwrite would be erasing something training rarely sampled.

## F What the overwrite turns from failing to passing in lane keeping

We roll out all 16,384 candidates of the 9,982 forgiven LK columns and of the 86,107 unforgiven mixed ones with the simulator that `pdm_score_full_v2` calls, and the human reference exactly as the forgiveness trigger does, reading positions at the bounding-box center that the lane-keeping check uses. The check recomputed from these rollouts equals the stored pre-overwrite column on all 1,574,322,176 candidates; it fails the human on every forgiven column and passes it on every unforgiven one. Table 5 compares the 124,222,171 flipped entries of the 9,831 mixed forgiven columns with the failing entries of the unforgiven ones. On the flipped entries, the largest distance from the route centerline outside intersections has median 6.09 m, and 84.7% of them reach further from it than the human rollout of the same scene, whose own largest distance has median 1.43 m over these columns. Their longest run above the 0.5 m limit spans a median of 33 of the rollout's 41 steps; the check fails a candidate at a run of 20. Rerun with the human rollout's path as its reference line, the same check fails 64.0% of the flipped entries, and 53.0% of all flipped entries fail it together with drivable-area compliance; 9.1% stay within 0.5 m of the human rollout laterally at every step. Of the flipped entries, 65.7% also fail drivable-area compliance, against 78.1% of the failing entries in the unforgiven columns. The human rollout's own largest distance from the route centerline is at most 1.0 m on 31.6% of the mixed forgiven columns, between 1.0 and 1.75 m on 27.8%, between 1.75 and 3.5 m on 31.5%, and above 3.5 m on 9.1%.

## G Reports of the same symptom on the upstream tracker

Users of the benchmark have reported machine-dependent scoring since 2024; the threads below were read on 21 September 2026. In `navsim` issue #28 (2024-10), the reporter writes "I changed a machine to run the evaluation, then the PDMS is 0.96, and every metric item is normal", and localizes it to

Table 5: Flipped entries of the mixed forgiven LK columns against the failing entries of the mixed unforgiven ones. "Against the human path" reruns the lane-keeping check with the human rollout's path as the reference line; "within 0.5 m" bounds the offset from the human rollout, perpendicular to its heading, at every 0.1 s step; the run is the check's longest run above 0.5 m outside intersections, of 41 steps. "Another check" is any of DAC, DDC, NC, TTC and TLC, read before the overwrite and after $b_j$. Percentages are over entries.

| | Forgiven, flipped | Unforgiven, failing |
|---|---|---|
| Columns | 9,831 | 86,107 |
| Entries | 124,222,171 | 481,727,638 |
| Fails LK against the human path (%) | 64.0 | 88.1 |
| Within 0.5 m of the human rollout (%) | 9.1 | 1.3 |
| Fails DAC (%) | 65.7 | 78.1 |
| Fails another check (%) | 82.4 | 91.8 |
| Largest distance from centerline, median (m) | 6.09 | 7.96 |
| Longest run above 0.5 m, median (steps) | 33 | 26 |

"the bug comes from `np.linalg.pinv()`, which generate different output given the same input". A second reporter there resolved it "by changing to another computer". A third attributes it to "some micro kernel selection issues for openblas … when running codes on CPUs with newer arch like intel 8457C", and prescribes a NumPy upgrade together with "purge all metric_caches and rebuild them". That part number is a Sapphire Rapids SKU, the same generation as the Sapphire Rapids parts used here. Issue #141 (2025-08) reports a zero PDM score on one accelerator and reasonable scores on another; a second reporter there is on a "Gen4 Xeon Scalable CPU", also Sapphire Rapids, and a third traces the zero to the drivable-area column and reports that fixing seeds, enabling deterministic algorithms and disabling TF32 "did not change anything". The maintainer asked in 2025-09 whether #141 "related to #28"; the thread records no answer. Issue #136 concerns instability in the same `pinv` call, and issue #81 concerns a metric mismatch attributed to library versions. These threads compare scores rather than stored artifacts, report no bitwise comparison and no reference artifact to compare against, and do not separate the CPU from the accelerator. The prescription in #28 to rebuild every metric cache after a library change already assumes that the cached artifacts depend on the machine; what is absent is a measurement of how far.

## H Estimators and intervals of Section 6

Section 6 reports the difference of median AUCs under one-to-one matching, $-0.2859$ [$-0.2939$, $-0.2797$]; direct standardization over strata of the failing-candidate count gives $-0.3733$ [$-0.3787$, $-0.3679$].

**How the interval and the matched contrast are formed.** Each scene contributes one AUC per column; a trapezoid-rule computation of the same quantity agrees to within 0.00000763. Ranks are averaged over ties, which scores a tie one half, and the AUC is the rank sum $(\sum_{\text{positive ranks}} - n_1(n_1 + 1)/2)/(n_1 n_0)$. AP and the partial AUC take the candidates in `torch.argsort` order with ties not grouped. On a rerun of the forward pass, grouping tied candidates, or breaking ties at random in each of 100 draws, leaves the median and the fifth and twenty-fifth percentiles of both, and the share of zero partial AUCs, unchanged at five decimals for the released model and each continued one, and every treatment finds the same scenes with a zero partial AUC; for the released model they are the 2,533 scenes behind Section 6's "at least a quarter". Matching is one-to-one on the failing-candidate count, without replacement, inside a caliper of 250 candidates. Treatment scenes take their partner in order of increasing supply, so the scenes with fewest admissible partners choose first; a treatment scene with no free partner inside the caliper is returned unmatched and counted. Balance after matching is read as the absolute standardized difference against a threshold of 0.1 and comes out at 0.016. The drivable-area column is mixed on all 9,831 forgiven scenes and on every matched partner, so all 7,940 pairs enter $D$. Direct standardization keeps every scene on both sides instead: strata are 500 candidates wide on the same count, each stratum is weighted by its number of treatment scenes, and every stratum holding treatment scenes also holds control scenes. The two estimators therefore read different forgiven scenes: the 7,940 matched ones have a median lane-keeping AUC of 0.7046, and

with the 1,891 that matching leaves out the median over all 9,831 falls to 0.6538. The contrast is a difference of medians, the median over the treatment group minus the median over the control group, not the median of per-pair differences. Its interval is a percentile bootstrap with 10,000 resamples at the 2.5 and 97.5 percentiles, drawn over pair indices with the same indices used on both groups so that a scene stays with its match; the matching itself is formed once and is not redone inside the resampling. The seed is given in the supplementary material.

**The Hydra-MDP checkpoint.** Table 6 reads the Hydra-MDP checkpoint released with GTRS with the scenes and pairs of Section 6 and the analyzer of Table 9 on navtrain and those of the held-out probe of Appendix K on navtest, beside the released GTRS-Dense checkpoint: R1 is the lane-keeping head's median AUC on the forgiven scenes, R2 the same on the matched unforgiven scenes and R3 the primary contrast $D_{\text{forgiven}} - D_{\text{matched}}$.

Table 6: Section 6's readings for the Hydra-MDP checkpoint released with GTRS and for the released GTRS-Dense checkpoint, with 95% intervals.

| Probe | Reading | Hydra-MDP | GTRS-Dense |
|---|---|---|---|
| navtrain | R1 | 0.6060 [0.5958, 0.6144] | 0.6538 [0.6465, 0.6609] |
| | R2 | 0.9891 [0.9888, 0.9895] | 0.9905 [0.9901, 0.9908] |
| | R3 | 0.3310 [0.3241, 0.3390] | 0.2850 [0.2770, 0.2922] |
| navtest | R1 | 0.6627 [0.6276, 0.6878] | 0.7095 [0.6847, 0.7241] |
| | R2 | 0.9761 [0.9732, 0.9784] | 0.9807 [0.9780, 0.9820] |
| | R3 | 0.2571 [0.2362, 0.2829] | 0.2368 [0.2206, 0.2544] |

## I Sources of the two devices in Section 6

Section 6 uses two established devices.

**Reading a main result against a control channel.** Hewitt & Liang (2019, Section 2) pair a linguistic task with a control task that assigns each word type a random output, and read the gap between the two accuracies (Hewitt & Liang, 2019, Section 1, Figure 2), which puts task accuracy in context with "the probe's capacity to memorize from word types" (Hewitt & Liang, 2019, abstract). The control there is a random relabeling and the quantity it bounds is a fitted probe's memorization. The control channel in Section 6 is a second output head of the same forward pass, scored against the same labels ($\text{AUC}_{\text{DAC}\to\text{LK}}$), and what it measures is the order a column retains without the head whose supervision was overwritten; in $D$ each head is instead scored against its own column ($\text{AUC}_{\text{DAC}\to\text{DAC}}$ and $\text{AUC}_{\text{LK}\to\text{LK}}$).

**Withholding an input to see what the score still returns.** Three audits of driving benchmarks remove an input and report what the score retains. Wei et al. (2026, Section 4.1) run two actor-blind probes on navtest and record EPDMS scores of 79.6 and 79.2 against human replay at 74.0. Li et al. (2023, Section 1) reproduce a planner that "relies solely on the ego status" and find it on par with UniAD and VAD on L2 distance and collision rate (Li et al., 2023, Section 4, "Ego status plays a key role"). Löwens et al. (2026, Section 1) replace a planner's camera input with memories retrieved from earlier traversals, leaving it able to "only access static and quasi-static information", and report that this is nearly sufficient on NAVSIM. Each of the three bounds what a benchmark score requires as input. The reading in Section 6 holds the input fixed and varies the head instead: what it compares is how much of one column's surviving order another head of the same pass already carries.

## J Order statistics and the seven-column share

**Order statistics and the median.** For each loss column, the number of candidates failing it before the overwrite is counted once per forgiven token-column. The $m$ counts are sorted ascending and read at five fixed ranks, indices 0, $\lfloor m/4 \rfloor$, $\lfloor m/2 \rfloor$, $\lfloor 3m/4 \rfloor$ and $m-1$. The median column of Table 2 is the value at rank $\lfloor m/2 \rfloor$, for even $m$ the upper of the two central values. The pool behind each row is every forgiven token-column of that column, the all-zero and the mixed together, which is why a row can reach the vocabulary width exactly.

**The seven-column share, column by column.** The numerator counts target entries over the seven loss columns whose written value the overwrite changed; the denominator is $103{,}288 \times 16{,}384 \times 7 = 11{,}845{,}894{,}144$. In a forgiven column the overwrite changes every entry that was not already passing: all 16,384 in a column constant at zero after $b_j$, and one per failing candidate in a mixed column. DAC and EP have no forgiven column (Table 2). Per column:

| Column | Changed entries | Counts at the five ranks |
|---|---|---|
| LK | 126,696,155 | 306, 10,484, 14,391, 15,604, 16,384 |
| TLC | 29,606,430 | 650, 12,365, 16,380, 16,384, 16,384 |
| DDC | 13,319,450 | 5,919, 14,857, 16,317, 16,384, 16,384 |
| TTC | 268,967 | 5,726, 15,384, 16,384, 16,384, 16,384 |
| NC | 158,216 | 11,343, 16,191, 16,384, 16,384, 16,384 |

The five add to 170,049,218, 1.4355% of the denominator; the share is the full quotient rounded to the nearest at four decimals.

**DDC before and after target mapping.** The 876 forgiven DDC columns partition into 490 three-valued, 373 two-valued and 13 one-valued; the raw and loss-effective mixed-column counts in Section 5.2 are cells of that partition, $490 + 373$ and 490, and the 386 columns in which no candidate scored 1.0 are the 373 and the 13 together. Under the mapping the 490 retain both binary labels and the 386 become all zero; the raw uniform columns are the 13.

## K Continued-training comparison

**Supervision and scoring.** L0, the positive control, continues at seed 0 on the published targets with every lane-keeping target set to passing. M averages each component's BCE over retained candidate entries of its unmasked tokens, with unchanged component weights and zero contribution if none remain in a batch. The imitation term and vocabulary dropout are unchanged; within each seed the arms share sample order and dropout random-number schedules. Checkpoints are selected by optimizer step independently of evaluation scores. Each model's stored trajectories are scored twice: OFF disables forgiveness; ON rebuilds the multiplicative and weighted aggregate inputs after human-zero overwrites, following the upstream correction cited in Section 8. On the released model, GTRS's own scoring with forgiveness on returns the OFF aggregate with the ON components, so ON rebuilds the aggregate as the upstream correction does. EC is identical under both switches and uses 10,040 adjacent-frame tokens: 1,428 in F and 8,612 in R (Section 8).

**Pairing, intervals and decision rule.** We pair by token and seed, bootstrap the 136 logs 10,000 times with replacement and shared draws for both subsets and all contrasts, and take percentile 95% intervals of token-weighted means, using only finite EC values. For each metric, switch and population, write $\Delta_s^X = \bar{q}_{X_s} - \bar{q}_{A_s}$ and $I_s^X = \Delta_s^X[F] - \Delta_s^X[R]$, with $X \in \{P, M\}$. A contrast is observed (O) if every seed has the same nonzero sign, every interval excludes zero, and the absolute seed-mean difference exceeds the within-arm seed scale $V_S(q) = \max_{X \in \{A,M,P\},\, i,j \in S} |\bar{q}_{X_i} - \bar{q}_{X_j}|$. For interactions, apply the same rule to $I_s^X$, replacing $\bar{q}_{X_s}$ in the scale by $D_s^X = \bar{q}_{X_s}[F] - \bar{q}_{X_s}[R]$. When the full rule fails, the contrast is unresolved: S (on the seed scale) if every interval excludes zero, U otherwise. Verdicts give $\mathcal{S} = \{0, 1\}$ followed by $\mathcal{S} = \{0, 1, 2\}$; the primary contrast is all-token OFF EPDMS for P–A. Point estimates round to nearest and interval endpoints outward at three decimals; scores are percentages and differences are score points, with no simultaneous-coverage adjustment.

Interactions: Table 4 gives O/O; all other $I^P$ and all $I^M$ of the scores in Table 8 are U/U.

Table 8 lists the EPDMS, LK and EC contrasts; DDC, TLC, NC, DAC, EP, TTC and HC do not separate either P–A or M–A under any scoring switch, subset or seed reading.

The positive control's OFF LK interval is below zero and its OFF EPDMS interval spans zero; continuation under A has an EPDMS interval spanning zero under both switches and a positive EC interval (Table 7).

All-token seed scales (two/three; column order of Table 7): 0.231/0.280; 0.225/0.251; 0.115/0.140; 0.214/0.280; 1.205/1.205.

Table 7: Absolute scores on all evaluated tokens, with EC restricted to adjacent-frame tokens; final rows give the positive-control and continuation differences [intervals].

| Arm | EPDMS OFF | EPDMS ON | LK OFF | LK ON | EC |
|---|---|---|---|---|---|
| Released | 86.560 | 90.038 | 85.790 | 96.460 | 75.926 |
| $A_0$ | 86.283 | 89.842 | 86.102 | 96.583 | 79.562 |
| $M_0$ | 86.456 | 89.889 | 86.407 | 96.830 | 79.253 |
| $P_0$ | 86.039 | 89.514 | 87.033 | 97.192 | 76.853 |
| $A_1$ | 86.480 | 89.988 | 86.218 | 96.666 | 78.357 |
| $M_1$ | 86.446 | 89.973 | 86.448 | 96.830 | 78.317 |
| $P_1$ | 86.270 | 89.739 | 86.983 | 96.978 | 76.773 |
| $A_2$ | 86.218 | 89.747 | 86.078 | 96.715 | 78.934 |
| $M_2$ | 86.362 | 89.886 | 86.358 | 96.871 | 78.347 |
| $P_2$ | 86.319 | 89.765 | 87.115 | 97.258 | 76.265 |
| $L_0$ | 86.237 | 89.770 | 84.884 | 95.612 | 79.104 |
| L0–A0 | -0.047 [-0.311, +0.237] | -0.072 [-0.323, +0.195] | -1.219 [-1.561, -0.872] | -0.972 [-1.259, -0.684] | -0.458 [-0.953, +0.049] |
| A0–Released | -0.277 [-0.743, +0.151] | -0.195 [-0.658, +0.224] | +0.313 [-0.019, +0.648] | +0.123 [-0.150, +0.396] | +3.635 [+2.842, +4.498] |

Table 8: Paired contrasts: point estimates [intervals] and two-/three-seed verdicts. Populations: all tokens, F and R as in Section 8; EC uses adjacent-frame tokens and is shared by both switches. R's per-token scores are the same under both switches.

| Contrast | Metric | Mode | Set | $s = 0$ | $s = 1$ | $s = 2$ | Verdict |
|---|---|---|---|---|---|---|---|
| P–A | EPDMS | OFF | All | -0.245 [-0.526, +0.044] | -0.210 [-0.431, +0.012] | +0.101 [-0.141, +0.357] | U/U |
| P–A | EPDMS | OFF | F | -0.167 [-0.752, +0.496] | -0.237 [-0.896, +0.401] | +0.416 [-0.290, +1.163] | U/U |
| P–A | EPDMS | OFF | R | -0.259 [-0.562, +0.049] | -0.205 [-0.444, +0.025] | +0.045 [-0.212, +0.313] | U/U |
| P–A | EPDMS | ON | All | -0.328 [-0.616, -0.031] | -0.249 [-0.473, -0.025] | +0.018 [-0.235, +0.281] | O/U |
| P–A | EPDMS | ON | F | -0.723 [-1.404, +0.008] | -0.497 [-1.208, +0.203] | -0.139 [-0.892, +0.648] | U/U |
| P–A | EPDMS | ON | R | -0.259 [-0.562, +0.049] | -0.205 [-0.444, +0.025] | +0.045 [-0.212, +0.313] | U/U |
| P–A | LK | OFF | All | +0.930 [+0.629, +1.256] | +0.766 [+0.467, +1.095] | +1.037 [+0.724, +1.381] | O/O |
| P–A | LK | OFF | F | +2.478 [+1.304, +3.758] | +3.524 [+2.008, +5.107] | +3.414 [+2.131, +4.694] | O/O |
| P–A | LK | OFF | R | +0.658 [+0.383, +0.944] | +0.281 [+0.060, +0.505] | +0.620 [+0.322, +0.983] | O/O |
| P–A | LK | ON | All | +0.609 [+0.370, +0.851] | +0.313 [+0.111, +0.513] | +0.543 [+0.286, +0.853] | O/O |
| P–A | LK | ON | F | +0.330 [+0.064, +0.624] | +0.496 [+0.169, +0.875] | +0.110 [-0.142, +0.381] | O/U |
| P–A | LK | ON | R | +0.658 [+0.383, +0.944] | +0.281 [+0.060, +0.505] | +0.620 [+0.322, +0.983] | O/O |
| P–A | EC | Both | All | -2.709 [-3.476, -1.943] | -1.584 [-2.329, -0.864] | -2.669 [-3.373, -2.007] | O/O |
| P–A | EC | Both | F | -7.633 [-9.986, -5.288] | -6.373 [-8.834, -3.911] | -6.303 [-8.842, -3.776] | O/O |
| P–A | EC | Both | R | -1.893 [-2.645, -1.135] | -0.790 [-1.448, -0.149] | -2.067 [-2.717, -1.458] | O/O |
| M–A | EPDMS | OFF | All | +0.173 [-0.031, +0.389] | -0.034 [-0.238, +0.177] | +0.144 [-0.039, +0.343] | U/U |
| M–A | EPDMS | OFF | F | +0.350 [-0.141, +0.872] | -0.037 [-0.470, +0.432] | -0.151 [-0.677, +0.343] | U/U |
| M–A | EPDMS | OFF | R | +0.142 [-0.076, +0.371] | -0.033 [-0.237, +0.173] | +0.196 [-0.014, +0.420] | U/U |
| M–A | EPDMS | ON | All | +0.047 [-0.158, +0.263] | -0.015 [-0.223, +0.198] | +0.139 [-0.051, +0.345] | U/U |
| M–A | EPDMS | ON | F | -0.493 [-1.144, +0.129] | +0.090 [-0.442, +0.658] | -0.186 [-0.724, +0.341] | U/U |
| M–A | EPDMS | ON | R | +0.142 [-0.076, +0.371] | -0.033 [-0.237, +0.173] | +0.196 [-0.014, +0.420] | U/U |
| M–A | LK | OFF | All | +0.305 [+0.083, +0.541] | +0.231 [+0.042, +0.435] | +0.280 [+0.064, +0.500] | O/O |
| M–A | LK | OFF | F | +0.551 [-0.138, +1.272] | +0.661 [-0.119, +1.463] | +0.771 [+0.000, +1.532] | U/U |
| M–A | LK | OFF | R | +0.261 [+0.043, +0.493] | +0.155 [-0.030, +0.354] | +0.194 [-0.010, +0.407] | U/U |
| M–A | LK | ON | All | +0.247 [+0.049, +0.453] | +0.165 [+0.008, +0.335] | +0.156 [-0.025, +0.349] | S/U |
| M–A | LK | ON | F | +0.165 [-0.095, +0.452] | +0.220 [+0.048, +0.460] | -0.055 [-0.240, +0.128] | U/U |
| M–A | LK | ON | R | +0.261 [+0.043, +0.493] | +0.155 [-0.030, +0.354] | +0.194 [-0.010, +0.407] | U/U |
| M–A | EC | Both | All | -0.309 [-0.860, +0.220] | -0.040 [-0.594, +0.482] | -0.588 [-1.111, -0.042] | U/U |
| M–A | EC | Both | F | -1.821 [-3.773, +0.065] | -1.120 [-3.253, +0.928] | -0.070 [-1.488, +1.427] | U/U |
| M–A | EC | Both | R | -0.058 [-0.608, +0.471] | +0.139 [-0.392, +0.665] | -0.673 [-1.193, -0.129] | U/U |

**Section 6's probe on the continued models.** Table 9 reruns the probe of Section 6 on the step-300 checkpoints over the same scenes and matched pairs: R1 is the lane-keeping head's median AUC on the 9,831 forgiven scenes, R2 the same on the 7,940 matched unforgiven scenes and R3 Section 6's primary contrast $D_{\text{forgiven}} - D_{\text{matched}}$, with verdicts by the rule above and L0–A0 read against both seed scales. On the matched unforgiven scenes every A, M and P model lies between 0.9919 and 0.9927, against 0.9905 for the released model.

Table 9: The probe of Section 6 on the continued models: point estimates of R1, R2 and R3, and two-/three-seed verdicts.

| Model | R1 | R2 | R3 |
|---|---|---|---|
| Released | 0.6538 | 0.9905 | 0.2850 |
| $A_0$ | 0.6257 | 0.9922 | 0.3225 |
| $M_0$ | 0.6962 | 0.9925 | 0.2378 |
| $P_0$ | 0.9225 | 0.9920 | 0.0570 |
| $A_1$ | 0.6423 | 0.9925 | 0.3007 |
| $M_1$ | 0.7044 | 0.9927 | 0.2239 |
| $P_1$ | 0.9233 | 0.9921 | 0.0563 |
| $A_2$ | 0.6141 | 0.9919 | 0.3219 |
| $M_2$ | 0.6965 | 0.9923 | 0.2363 |
| $P_2$ | 0.9230 | 0.9919 | 0.0567 |
| $L_0$ | 0.4447 | 0.3003 | -0.1318 |
| P–A | O/O | O/U | O/O |
| M–A | O/O | S/S | O/O |
| L0–A0 | O/O | O/O | O/O |

**Section 6's probe on held-out navtest.** Table 10 repeats the probe on the 1,496 navtest scenes whose human lane-keeping score is zero and whose lane-keeping column is mixed, and on 1,287 matched pairs, with R4 the matched minus the forgiven median over the pairs, the sign opposite to Section 6's difference of medians. P–A is observed on every reading and P's R3 and R4 stay above zero in every seed; M–A is observed only on R4, and L0's R1 and R2 lie below 0.5.

Table 10: The probe of Section 6 on held-out navtest scenes: point estimates of R1 to R4 and two-/three-seed verdicts.

| Model | R1 | R2 | R3 | R4 |
|---|---|---|---|---|
| Released | 0.7095 | 0.9807 | 0.2368 | 0.2441 |
| $A_0$ | 0.7150 | 0.9781 | 0.2268 | 0.2327 |
| $M_0$ | 0.7308 | 0.9815 | 0.2126 | 0.2190 |
| $P_0$ | 0.8236 | 0.9834 | 0.1399 | 0.1421 |
| $A_1$ | 0.7184 | 0.9813 | 0.2249 | 0.2285 |
| $M_1$ | 0.7293 | 0.9833 | 0.2138 | 0.2180 |
| $P_1$ | 0.8397 | 0.9850 | 0.1234 | 0.1327 |
| $A_2$ | 0.7105 | 0.9788 | 0.2341 | 0.2380 |
| $M_2$ | 0.7204 | 0.9818 | 0.2170 | 0.2249 |
| $P_2$ | 0.8356 | 0.9830 | 0.1275 | 0.1323 |
| $L_0$ | 0.4778 | 0.2426 | -0.2291 | -0.2192 |
| P–A | O/O | O/O | O/O | O/O |
| M–A | U/U | S/S | S/S | O/O |
| L0–A0 | O/O | O/O | O/O | O/O |

**Distance to the human trajectory.** Table 11 compares each model's navtest plans with the human trajectory at eight shared times, with the pairing, intervals and rule above.

**Step 530.** Table 12 reads the same ten runs at optimizer step 530 with the pairing, intervals and rule above; L0–A0's EPDMS intervals lie below zero under both switches.

**The probes at step 530.** Table 13 repeats both probes on the same scenes and pairs at step 530, and P–A keeps its step-300 sign and verdict on R1 and R3 of both. On navtest, P's R1 exceeds A's by 0.1264–0.1377, and P's R3 is 0.1100–0.1156 against 0.2171–0.2271 for A.

Table 11: Planned-to-human distance in meters, $P_s$ minus $A_s$ on F, on R and their difference: point estimates [intervals] and two-/three-seed verdicts. ADE averages the eight aligned poses, FDE takes the last, and the lateral deviation is the largest distance from a planned pose to the human path.

| Metric | Set | $s = 0$ | $s = 1$ | $s = 2$ | Verdict |
|---|---|---|---|---|---|
| ADE | F | +0.084 [+0.062, +0.108] | +0.054 [+0.035, +0.074] | +0.050 [+0.030, +0.070] | O/O |
| ADE | R | +0.034 [+0.023, +0.048] | +0.007 [-0.003, +0.017] | +0.023 [+0.009, +0.041] | U/U |
| ADE | F−R | +0.049 [+0.024, +0.075] | +0.047 [+0.026, +0.070] | +0.027 [+0.001, +0.050] | O/O |
| FDE | F | +0.201 [+0.142, +0.267] | +0.123 [+0.075, +0.175] | +0.090 [+0.038, +0.141] | O/O |
| FDE | R | +0.095 [+0.068, +0.125] | +0.018 [-0.005, +0.041] | +0.036 [+0.004, +0.074] | U/U |
| FDE | F−R | +0.106 [+0.041, +0.177] | +0.105 [+0.051, +0.161] | +0.055 [-0.009, +0.113] | O/U |
| Lateral | F | +0.124 [+0.086, +0.166] | +0.131 [+0.092, +0.168] | +0.039 [-0.001, +0.078] | O/U |
| Lateral | R | +0.036 [+0.024, +0.050] | +0.014 [-0.002, +0.029] | -0.041 [-0.053, -0.028] | U/U |
| Lateral | F−R | +0.087 [+0.047, +0.131] | +0.117 [+0.076, +0.157] | +0.080 [+0.037, +0.121] | O/O |

Table 12: Step 530: $P_s$ and $M_s$ minus $A_s$ on all tokens, F and R, and the interactions $I_s^X$ (F−R), as point estimates with two-/three-seed verdicts (V); the last rows give L0–A0 [intervals].

| | | | P–A | | | | M–A | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Mode | Set | $s = 0$ | $s = 1$ | $s = 2$ | V | $s = 0$ | $s = 1$ | $s = 2$ | V |
| EPDMS | OFF | All | -0.259 | +0.064 | -0.302 | U/U | +0.075 | -0.218 | +0.059 | U/U |
| | | F | +0.212 | +0.193 | -0.500 | U/U | +0.332 | -0.123 | -0.184 | U/U |
| | | R | -0.341 | +0.041 | -0.267 | U/U | +0.030 | -0.235 | +0.102 | U/U |
| | | F−R | +0.553 | +0.152 | -0.234 | U/U | +0.303 | +0.111 | -0.285 | U/U |
| EPDMS | ON | All | -0.333 | -0.051 | -0.365 | U/U | +0.072 | -0.266 | +0.020 | U/U |
| | | F | -0.287 | -0.576 | -0.924 | U/U | +0.314 | -0.448 | -0.446 | U/U |
| | | R | -0.341 | +0.041 | -0.267 | U/U | +0.030 | -0.235 | +0.102 | U/U |
| | | F−R | +0.054 | -0.617 | -0.657 | U/U | +0.284 | -0.213 | -0.548 | U/U |
| LK | OFF | All | +1.194 | +1.021 | +1.087 | O/O | +0.140 | -0.016 | +0.403 | U/U |
| | | F | +3.634 | +3.579 | +2.974 | O/O | +0.055 | +0.275 | +0.551 | U/U |
| | | R | +0.765 | +0.571 | +0.755 | O/O | +0.155 | -0.068 | +0.378 | U/U |
| | | F−R | +2.870 | +3.008 | +2.218 | O/O | -0.100 | +0.343 | +0.173 | U/U |
| LK | ON | All | +0.634 | +0.494 | +0.626 | O/O | +0.123 | -0.074 | +0.321 | U/U |
| | | F | -0.110 | +0.055 | -0.110 | U/U | -0.055 | -0.110 | +0.000 | U/U |
| | | R | +0.765 | +0.571 | +0.755 | O/O | +0.155 | -0.068 | +0.378 | U/U |
| | | F−R | -0.875 | -0.516 | -0.865 | O/O | -0.210 | -0.042 | -0.378 | U/U |
| EC | Both | All | -2.321 | -1.285 | -2.291 | O/O | -0.060 | -0.010 | -0.528 | U/U |
| | | F | -7.423 | -5.672 | -8.123 | O/O | -0.980 | +0.210 | -1.821 | U/U |
| | | R | -1.475 | -0.557 | -1.324 | U/U | +0.093 | -0.046 | -0.314 | U/U |
| | | F−R | -5.948 | -5.115 | -6.800 | O/O | -1.073 | +0.257 | -1.507 | U/U |
| EPDMS | OFF | L0–A0 | -0.256 [-0.470, -0.051] | | | | | | | |
| EPDMS | ON | L0–A0 | -0.265 [-0.494, -0.044] | | | | | | | |
| LK | OFF | L0–A0 | -1.169 [-1.516, -0.835] | | | | | | | |
| LK | ON | L0–A0 | -1.021 [-1.318, -0.731] | | | | | | | |
| EC | Both | L0–A0 | -0.349 [-1.011, +0.302] | | | | | | | |

Table 13: The probes of Section 6 at step 530 for the A, M and P models: point estimates on navtrain (R1 to R3) and on held-out navtest (R1 to R4), and two-/three-seed verdicts.

| | navtrain | | | navtest | | | |
|---|---|---|---|---|---|---|---|
| Model | R1 | R2 | R3 | R1 | R2 | R3 | R4 |
| $A_0$ | 0.6316 | 0.9928 | 0.3092 | 0.7208 | 0.9815 | 0.2171 | 0.2277 |
| $M_0$ | 0.7231 | 0.9932 | 0.2083 | 0.7200 | 0.9833 | 0.2145 | 0.2224 |
| $P_0$ | 0.9419 | 0.9929 | 0.0416 | 0.8531 | 0.9852 | 0.1129 | 0.1166 |
| $A_1$ | 0.6164 | 0.9924 | 0.3275 | 0.7302 | 0.9801 | 0.2245 | 0.2267 |
| $M_1$ | 0.7010 | 0.9931 | 0.2326 | 0.7331 | 0.9832 | 0.2151 | 0.2157 |
| $P_1$ | 0.9425 | 0.9927 | 0.0417 | 0.8566 | 0.9854 | 0.1100 | 0.1136 |
| $A_2$ | 0.6359 | 0.9928 | 0.3060 | 0.7101 | 0.9814 | 0.2271 | 0.2372 |
| $M_2$ | 0.7193 | 0.9928 | 0.2145 | 0.7287 | 0.9831 | 0.2114 | 0.2196 |
| $P_2$ | 0.9393 | 0.9924 | 0.0437 | 0.8478 | 0.9848 | 0.1156 | 0.1212 |
| P–A | O/O | U/U | O/O | O/O | O/O | O/O | O/O |
| M–A | O/O | O/U | O/O | U/U | O/O | U/U | U/U |

## L Solver sensitivity of the human census

The census uses `np.linalg.pinv`; Wei et al. (2026, solver matrix in Section 4.2) report human-reference failure rates of 100.0%, 12.9%, and 9.1% on their own 450-scene interactive pool for stock pseudoinverse, direct solve, and Hermitian pseudoinverse, alongside rollout-divergence rates of 99.8%, 0%, and 0%. Under the software configuration they audit (supplementary material), substituting `scipy.linalg.pinvh` or direct solve leaves all 103,288 tokens' nine human values bitwise identical, preserves every per-column human-zero set and their union, and retains agreement on all 1,000 worst-conditioned tokens, although the per-token relative matrix difference between stock and Hermitian pseudoinverses reaches $1.159 \times 10^{-12}$. Their fallback definition uses deviation from the proposal and single-step distance, while their threshold-ladder description says "path and step limits" (Wei et al., 2026, Sections 3.2 and 4.3); we report both readings. Under the path-length reading, the 95th-percentile path is 40.67 m, so at least 5,165 rollouts exceed the strict 20 m limit, and none exceeds 100 m, the longest path being 65.85 m. Both readings detect their kilometer-scale failure, a median path length of 32.11 km in the affected configuration of their 32-token diagnostic, whereas the choice of quantity changes the strict-threshold conclusion on our corpus. Their overwrite rates are measured in the divergent condition their own solver table reports; under the deviation-or-step reading, the 103,288 stock-solver human rollouts of this split record 0 exceedances at $(20, 10)$, $(100, 50)$, and $(200, 100)$ m, with every rollout within 5.53 m of its proposal and no single step above 2.27 m, and forgiveness still erases a distinction in a loss column on 10.8793% of the same scenes.