BadTouch: The First Study of Data Poisoning in Tactile Models

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated tactile observations share a physical contact, but their supplied force labels need not share the same quality. How should a curator use this provenance when adapting a frozen tactile encoder? We study label-only poisoning on 904 public GelSight trajectories with same-input controls, two frozen encoder settings, and explicit verified-label budgets. A finite-bank deletion frontier separates errors made by a detector from correct samples necessarily lost through whole-contact removal. In the Sparsh curation bank, complete poison removal by whole-contact deletion requires deleting 93.69% of correct sharp candidate rows. Using the same trusted predictor and contact-calibrated thresholds, row deletion reduces this loss to 0.26% and sharp MAE from 0.092 to 0.054 N, at 99.67% poison recall for the 1.5 N target. This collateral improvement does not establish additional task value: when all 362 verified contacts can be reused, excluding unverified sharp candidates reaches ordinary/sharp MAE of 0.025/0.040 N, compared with 0.026/0.039 N after filtered augmentation. Lower-budget sharp gains vary across seeds. Source-allocation, calibration-scarcity, accepted-label and recovery controls delimit these findings within one acquisition population. The results show why provenance-aware curation must evaluate both deletion cost and predictive utility under the same available verified information.

1 INTRODUCTION

Frozen tactile encoders make force adaptation inexpensive: a reusable representation converts contact images into features, and a small downstream model learns the new task (Higuera et al., 2025; Feng et al., 2025; 2026). This division of labor also creates a specific trust boundary. Keeping the encoder unchanged does not prevent an untrusted data contributor from associating a distinctive contact condition with an incorrect force label. The relevant question is which assumptions sustain that association, and which measurements distinguish it from ordinary prediction error.

Repeated observations create a second trust boundary: a physical contact can supply both correct and incorrect force labels. Grouping all its observations is essential for data splitting and calibration, but deleting the entire contact can discard useful labels. A successful screen must therefore be evaluated through the action it triggers and the predictor trained afterward. That final comparison must also account for labels the curator has already verified: retaining untrusted data need not outperform training directly on the available trusted information.

We study this problem on public GelSight force data. Flat and spherical indenters define the ordinary task; a separately recorded sharp condition supplies candidates whose selected labels are replaced by 1.5 N or 5 N. The former lies inside the pooled observed range, while the latter is caught by a basic label audit. Same-input correct-label controls measure the effect of these interventions. We then separate the unavoidable cost of whole-contact deletion from detector error, compare row deletion under the same screen, and test whether the retained candidates add task value when all verified labels can be reused.

Contributions. We separate three questions in tactile data curation: what a label intervention changes, what a deletion policy necessarily discards, and whether retained data improve the final predictor. Our contributions are (i) matched-input poisoning experiments that account for contact-source allocation; (ii) an exact whole-contact collateral frontier and a contact-calibration corollary for

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.

row deletion; and (iii) retrained comparisons that include verified-label reuse, category exclusion and trusted recovery. Together these distinguish a gain in candidate retention from a gain in task utility. The artifact records contact partitions, supplied and true labels, filtering decisions and predictions for independent reconstruction of the comparisons.

2 SETTING AND MEASUREMENTS

Let $S \in \{\text{flat}, \text{sphere}, \text{sharp}\}$ denote the recorded indenter condition and G a source contact trajectory. In the primary Sparsh setting, the frozen encoder maps a current/previous tactile image pair X to $Z \in \mathbb{R}^{768}$; a head $g_\theta(Z)$ predicts normal force $Y = |F_z|$ in newtons. Appendix G.3 defines the second, single-frame 384-dimensional setting. The contributor can supply selected sharp training samples and assign their labels. It cannot change the encoder, victim optimizer, held-out inputs, or sensor signals. Condition, batch, and trajectory metadata are verified and outside contributor control. Sharp is a recorded condition, not a paired transformation of an otherwise identical flat/sphere contact. Geometry is therefore not causally separated from condition-specific acquisition and force statistics.

Two budgets. Let D_c contain n ordinary training examples, P the poisoned sharp indices, and B disjoint correctly labelled sharp indices. For target a,

$$D' = D_c \cup \{(X_i, a) : i \in P\} \cup \{(X_i, Y_i) : i \in B\}, \qquad p_f = \frac{|P|}{n + |P| + |B|}. \tag{1}$$

We report the number U_P of represented poison source trajectories, its fraction $c_T = U_P/U_T$ among available sharp training sources, and its share $s_P = U_P/[U_c + |G_P \cup G_B|]$ among all represented sources. A trajectory may contribute both correct and corrupted labels. These are provenance quantities, not counts of entirely corrupted events, independent samples, or physical acquisition costs.

Paired regression outcomes. For a per-frame quantity h_i, write $\langle h \rangle_T$ for its mean within each held-out sharp trajectory, averaged equally across trajectories. Our primary corruption outcome compares a poisoned head g_a with its all-correct, same-input reference g_0:

$$\Delta_T = \langle |g_a(Z) - Y| - |g_0(Z) - Y| \rangle_T. \tag{2}$$

We also report target error $\langle |g_a(Z) - a| \rangle_T$, target-band rate $\langle \mathbf{1}[|g_a(Z) - a| \le 0.1\,\text{N}] \rangle_T$ with its clean-reference rate, and the ≥ 4 N exceedance rate for the 5 N construction. Target-band membership alone is insufficient when legitimate force can already be near 1.5 N. Ordinary MAE uses the same trajectory weighting; ordinary R^2 pools frames and therefore has different weighting. Non-trigger false activation counts ordinary estimates ≥ 4 N. Every seed is retained; means and sample standard deviations describe five repeated runs on the same finite dataset, not independent sensors.

Range controls. A 3 N label audit separates every 5 N poison from the retained legitimate labels, but cannot identify a 1.5 N label replacement. Independently, an output cap $\bar{g} = \min(g, 3)$ makes $\bar{g} \ge 4$ impossible. That is an algebraic restriction on a metric, not proof of accurate force estimates or physical safety. We therefore evaluate both target-band response and paired true-label error, and measure downstream performance after label filtering.

Three comparisons. Verified labels denote the original sensor labels reserved as trusted in this experiment; no independent physical remeasurement is performed. Budgets count access to trusted contacts, rather than verification labor. Same-input label controls hold the training inputs and optimization setup fixed and identify the effect of the supplied labels. The deletion study holds the candidate bank, reference predictor and thresholds fixed, then compares complete filtering-and-retraining pipelines; their retained inputs, scalers and sampling exposure can differ. The utility study additionally lets every model reuse all labels already verified for reference fitting and calibration. Because corruption is confined to sharp candidates and condition metadata are trusted, this study includes excluding all unverified sharp rows while retaining verified sharp data.

3 CONTACT PROVENANCE AND CURATION

3.1 THE COLLATERAL COST OF WHOLE-CONTACT REMOVAL

Let contact j supply p_j poisoned and b_j correctly labelled sharp candidate rows, with $K = \sum_j p_j > 0$ and $B = \sum_j b_j > 0$. A whole-contact policy deletes every supplied row from each selected contact.

Proposition 1 (Collateral at complete poison removal)**.** *Every whole-contact policy with poison recall one loses at least*

$$L_{\text{contact}} = \frac{\sum_{j:p_j>0} b_j}{B} \tag{3}$$

of correctly labelled sharp candidate rows. Deleting exactly the poison-bearing contacts attains the bound.

The proof is immediate: every contact containing a poison must be removed. This oracle bound diagnoses a cost of the permitted action even when poison localization is perfect. It concerns candidate deletion, not the error of the retrained model. Appendix D gives the sampling expectation and the exact partial-recall frontier, computed by 0–1 knapsack with attaining subsets. We test whether changing the action to row deletion preserves useful data under the same calibrated screen.

3.2 CALIBRATE BY CONTACT, DELETE BY ROW

A calibration unit need not be a deletion unit. Let h be a regressor fitted on separate verified trajectories and $e_i = |\widetilde{Y}_i - h(Z_i)|$ the residual of a supplied label. Within each condition/batch stratum, calibrate the trajectory score $E_g = \max_{i \in g} |Y_i - h(Z_i)|$ on n_{cal} complete, verified contacts. Set q to order statistic $\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$, or $+\infty$ if that rank exceeds n_{cal}. At $\alpha = 0.05$, fewer than 19 calibration contacts in a stratum force $q = +\infty$; Appendix G.1 tests this calibration-budget constraint. We compare two actions using the *same* h and q:

$$D_{\text{row}} = \{i : e_i > q\}, \qquad D_{\text{contact}} = \bigcup_{g:\max_{i \in g} e_i > q} g. \tag{4}$$

Always $D_{\text{row}} \subseteq D_{\text{contact}}$, so row deletion cannot remove more correct rows. It can retain a poison whose own residual is small even when another row flags the contact; equal poison recall must be measured.

For a new parent contact exchangeable with calibration contacts within its condition/batch, the standard split-rank argument gives $\Pr\{E_g > q\} \leq \alpha$ (Lei et al., 2018). Under label-only corruption, let C_g be its rows whose supplied labels remain correct. Their residuals are unchanged, so $\{D_{\text{row}} \cap C_g \neq \varnothing\} \subseteq \{E_g > q\}$ and the same marginal bound applies. A corrupted row can nevertheless cause the whole-contact action to delete every correct row in its contact. This corollary protects correct rows under its exchangeability assumption; it does not guarantee poison rejection. Conditioning on attacker-selected parent contacts can invalidate that assumption, so candidate-bank collateral is measured empirically. Appendix D.1 supplies the proof and a separating example.

4 DATA AND EXECUTION PROTOCOL

We independently download the compact public GelSight images and force records, verify their release hashes, and retain up to 24 uniformly spaced loader examples per trajectory. The selected flat/sphere/sharp conditions in batches 1 and 2 give 904 trajectories and 20,580 image pairs. The public release contains additional sphere batches; our study uses the six stated condition/batch banks. Retained legitimate force ranges approximately from 0.063 to 1.80241 N. The documented collection protocol uses normal loads up to 3 N (AI at Meta, 2024); this is an acquisition envelope, not a certified hardware safety bound.

The primary representation uses the released frozen Sparsh DINO ViT-B with its native current/previous-frame preprocessing, producing 768 features per observation. The second setting uses a single current frame and a 384-dimensional ImageNet DINO representation. Complete contacts remain within one partition. Release hashes, background-reference checks and duplicate audits verify the selected data and exclude shared source-frame indices and exact-pixel duplicates

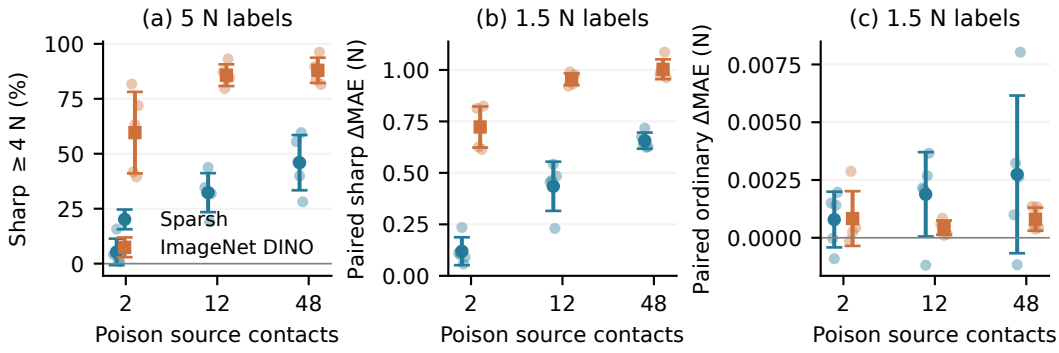

Figure 1: Fixed 48-frame poison count in two frozen encoder settings. All 2/12/48-source arms share source eligibility and per-batch counts. Dots are five runs; larger markers and bars are means and sample SD. MAE changes use same-input, correct-label controls within each encoder. Sparsh and ImageNet DINO differ in input, encoder and head capacity; their comparison assesses breadth, not a causal effect of pretraining.

across contacts. Appendix A gives pinned versions, preprocessing, input construction and the full inventory.

Primary runs split complete trajectories 60/20/20 into fitting, validation, and test partitions within every condition/batch, using seeds 0–4. The default head is $768 \to 512 \to 256 \to 1$ with GELU, AdamW at learning rate 10^{-3} and weight decay 10^{-4}, and uniformly sampled minibatches of 256. A clean ordinary-only baseline selects 1,000, 2,000, or 4,000 updates by ordinary validation MAE. Every paired arm uses that seed's selected budget and initialization; attack responses never select settings. Features are standardized using only the final observed training inputs. Same-input arms consequently share the scaler and minibatch sequence. Targets remain in newtons. All extraction and training are executed on CPU. All outcomes have retained per-run predictions; the second-encoder and trusted-data follow-ups use the protocols in Appendices G–I.

5 Controlled Empirical Results

5.1 Label Corruption and Contact-Source Allocation

Every arm corrupts 48 frames (approximately 0.58% of training inputs), drawn from the same eligible source pool with 24 frames from each batch. For the 1.5 N target, Sparsh sharp MAE increases over same-input correct-label controls by 0.119 ± 0.068, 0.435 ± 0.119 and 0.657 ± 0.039 N for 2, 12 and 48 source contacts. At 48 sources, ordinary MAE increases by 0.0027 ± 0.0034 N. Thus a label inside the pooled observed range can damage the sharp condition while leaving ordinary estimation nearly unchanged. Figure 1 also shows the 5 N construction and the second frozen encoder. In the latter, two sources already induce a 0.723 N paired sharp-MAE increase for the 1.5 N target. Source allocation changes the measured response, but broad source access is not required in every encoder setting. Tables 6–8 and Appendix G.3 report all matched controls and the separate uniform-sampler references.

5.2 The Cost of the Deletion Action

Curation uses a 20/20/40/20 contact split for trusted fitting, calibration, victim fitting and test. The victim receives ordinary contacts, approximately 5% poisoned sharp rows and equal correct sharp cover. Each retained bank trains the same MLP architecture for 2,000 updates from the same seed-specific initialization. Filtering changes its fitted scaler and sampling exposure. The Ridge reference uses all three conditions with fitting-only penalty selection and disjoint condition/batch calibration.

At 1.5 N, unfiltered sharp MAE is 0.491 ± 0.066 N versus 0.047 ± 0.005 N for the all-correct MLP. Whole-contact residual deletion removes every poison but loses $93.69 \pm 2.41\%$ of correct sharp

cover. In every seed this equals Equation 3: the detector reaches the complete-recall contact-deletion frontier, with zero excess sharp-cover deletion. Its resulting sharp MAE is 0.092 ± 0.012 N.

Table 1: Same residual predictor and contact-calibrated thresholds; only deletion granularity changes. Five seeds, mean $\pm$ sample SD. The all-correct MLP has sharp MAE 0.047 ± 0.005 N and ordinary MAE 0.034 ± 0.003 N.

Action, target	Poison recall (%)	Cover loss (%)	Sharp MAE (N)	Ord. MAE (N)
Contact, 1.5 N	100.00 ± 0.00	93.69 ± 2.41	0.092 ± 0.012	0.033 ± 0.003
Row, 1.5 N	99.67 ± 0.33	0.26 ± 0.27	0.054 ± 0.005	0.034 ± 0.001
Contact, 5 N	100.00 ± 0.00	93.69 ± 2.41	0.092 ± 0.012	0.033 ± 0.003
Row, 5 N	100.00 ± 0.00	0.26 ± 0.27	0.053 ± 0.006	0.033 ± 0.003

Using the same reference and thresholds, row deletion reduces cover loss to $0.26 \pm 0.27\%$ and sharp MAE to 0.0538 ± 0.0049 N, a paired reduction of 0.0379 ± 0.0088 N. It removes $99.67 \pm 0.33\%$ of 1.5 N poisons and all 5 N poisons. Both actions flag exactly the same contacts; surviving 1.5 N poisons have small individual residuals in otherwise flagged contacts. Thus contact-level calibration permits substantially lower candidate loss when deletion acts on rows. The second encoder shows the same deletion-cost pattern: contact/row cover loss is 93.69/0.33% at full recall, with sharp MAE 0.074/0.037 N (Appendix G.3).

These comparisons concern a fixed candidate bank. Feature-only novelty is a diagnostic control: changing a label on the same input leaves its score unchanged. Appendix E.4 reports paired annotation tests; Appendix F gives full curation contrasts, the deletion frontiers, and reference-fitting coverage controls.

5.3 Does lower collateral add task value?

The preceding comparison measures the consequences of deletion on a fixed candidate bank. A final learner can also reuse the labels already verified by the curator. We therefore compare verified-only training, verified plus ordinary candidates, and verified plus filtered candidates at total verified-contact budgets of 126, 162 and 362. The full budget combines 181 reference-fitting and 181 calibration contacts, including 122 sharp contacts; the lowest budget still includes 42 verified sharp contacts. All heads use the same architecture and 2,000 updates; every budget includes verified sharp contacts.

Adding ordinary candidates improves ordinary MAE in all 15 seed/budget pairs. At the full budget, excluding unverified sharp candidates reaches 0.0253/0.0402 N ordinary/sharp MAE, compared with 0.0259/0.0393 N after filtered augmentation. The ordinary-data gain therefore does not depend on retaining unverified sharp rows. Holding the filtered ordinary pool fixed, retained sharp augmentation changes sharp MAE by -0.0043 ± 0.0195, -0.0021 ± 0.0151 and -0.0014 ± 0.0040 N across the three budgets. These gains vary across seeds and do not establish consistent incremental sharp-task value. The comparisons include changes in scaling and sample exposure; Appendix J reports the paired contrasts, and Appendix I checks one count-derived alternative to the equal-update schedule.

5.4 Accepted labels and trusted recovery

A contributor that knows the fixed reference h and threshold q can choose labels within the residual acceptance interval and audited range. We retain the same 5% poison indices and equal correct sharp cover, and choose the largest float32 value in each nonempty interval

$$[\max(0, h(Z_i) - q),\ \min(3, h(Z_i) + q - 10^{-6})]. \tag{5}$$

No test observations or outcome search select these labels. All poisons survive row screening, and the resulting sharp-MAE excess over an all-correct head on exactly the retained inputs is 0.1545 ± 0.0390 N (Appendix F.3).

We fine-tune both predecessors on the same 181 verified fitting contacts for 2,000 further updates, retaining their scalers and restarting AdamW. This fixed schedule reduces the paired excess to 0.0075 ± 0.0039 N; the corrupted model's absolute sharp MAE falls from 0.202 to 0.053 N (Table 21). The construction separates correct-row protection from poison rejection, while paired recovery shows that screen evasion alone does not establish persistence after trusted adaptation.

Table 2: Equal verified-label access with a category-exclusion baseline. Ordinary augmentation discards all unverified sharp candidates and needs no residual screen. All heads use 2,000 updates; N, mean ± sample SD over five reused splits.

Verified	Victim training	Ordinary MAE	Sharp MAE
126	Verified only	0.048 ± 0.005	0.065 ± 0.007
126	Verified + ordinary	0.030 ± 0.003	0.060 ± 0.008
126	Verified + filtered candidates	0.030 ± 0.004	0.054 ± 0.021
162	Verified only	0.044 ± 0.003	0.060 ± 0.004
162	Verified + ordinary	0.029 ± 0.002	0.055 ± 0.006
162	Verified + filtered candidates	0.029 ± 0.003	0.053 ± 0.015
362	Verified only	0.031 ± 0.002	0.040 ± 0.006
362	Verified + ordinary	0.025 ± 0.002	0.040 ± 0.006
362	Verified + filtered candidates	0.026 ± 0.002	0.039 ± 0.006

Table 3: Label-aware follow-up with all 362 verified contacts reused by every final head. MAE and the paired sharp change from each method's own all-correct world are in N; mean ± sample SD across five reused splits. The boundary bank is transferred to the two BayesClean I1 adaptations; it is not adaptive to them.

Label world	Candidate treatment	Ordinary MAE	Sharp MAE	Sharp Δ
All correct	Unfiltered	0.026 ± 0.002	0.036 ± 0.006	0.000 ± 0.000
All correct	Original residual	0.026 ± 0.003	0.036 ± 0.007	0.000 ± 0.000
All correct	BayesClean I1 (std.)	0.027 ± 0.003	0.038 ± 0.007	0.000 ± 0.000
All correct	BayesClean I1 (raw)	0.025 ± 0.001	0.036 ± 0.007	0.000 ± 0.000
1.5 N	Unfiltered	0.036 ± 0.003	0.200 ± 0.030	0.164 ± 0.029
1.5 N	Original residual	0.026 ± 0.002	0.039 ± 0.006	0.003 ± 0.003
1.5 N	BayesClean I1 (std.)	0.027 ± 0.002	0.040 ± 0.005	0.002 ± 0.002
1.5 N	BayesClean I1 (raw)	0.027 ± 0.002	0.039 ± 0.006	0.003 ± 0.003
Boundary transfer	Unfiltered	0.031 ± 0.003	0.095 ± 0.016	0.059 ± 0.014
Boundary transfer	Original residual	0.031 ± 0.004	0.103 ± 0.012	0.067 ± 0.014
Boundary transfer	BayesClean I1 (std.)	0.029 ± 0.003	0.046 ± 0.007	0.009 ± 0.005
Boundary transfer	BayesClean I1 (raw)	0.026 ± 0.002	0.039 ± 0.005	0.004 ± 0.002

5.5 Label-aware sanitizer controls

At the same 362-contact budget, two BayesClean I1 adaptations share a candidate-fitted posterior but apply acceptance in standardized or original target coordinates. Each has an all-correct control. Supplied-candidate normalization differs from the source's clean-normalization setting; Appendix K specifies both rules and information access.

For 1.5 N replacements, sharp MAE is 0.0393 N after trusted residual screening and 0.0397/0.0386 N under standardized/raw I1 filtering. Their similar final errors conceal different candidate losses: the I1 variants remove 58.59%/6.29% of ordinary rows and 15.13%/0.20% of correct sharp cover, while every verified row remains. The frozen boundary labels from the previous subsection transfer less well to I1 filtering: sharp MAE is 0.1031 N for the known residual screen, versus 0.0464/0.0394 N for standardized/raw I1. Table 3 also gives each pipeline's own-clean excess. These end-to-end outcomes include filter membership changes; transfer from the residual-screen attack does not establish I1-adaptive robustness.

Normalizing with the same verified contacts changes this tradeoff (Appendix L): under 1.5 N replacements, raw I1 recall rises from 98.23% to 98.95%, but correct-cover loss rises from 0.20% to 2.70% and mean sharp MAE from 0.0386 to 0.0395 N. This does not establish a uniform utility improvement or remove coordinate dependence.

Sanitizer-aware labels. To remove the transferred-attack gap, a 20-query stress test updates fixed poison labels toward the verified-normalized raw-I1 boundary (Appendix M). A fresh posterior rejects $29.16 \pm 11.36\%$, versus $98.95 \pm 0.95\%$ for fixed 1.5 N labels. On identical retained inputs, correct

labels lower sharp MAE from 0.04053 to 0.03568 N: a 0.00485 ± 0.00170 N effect. Correct sharp-cover/ordinary losses are $25.13/24.44\%$. Every seed misses complete acceptance; this unconverged heuristic is not a worst-case attack.

5.6 Complete acquisition-batch holdout

With source-only validation, 5% poisoning and same-input controls, batch $1 \rightarrow 2$ reaches $50.5 \pm 8.7\%$ sharp exceedance but raises ordinary MAE from 0.059 to 0.889 N; the reverse direction has no target-band response. Target-batch background calibration is allowed. Neither direction establishes utility-preserving transfer (Tables 14 and 15).

6 Related work

Tactile learning and embodied attacks. GelSight and DIGIT provide visual contact measurements (Yuan et al., 2017; Lambeta et al., 2020). Sparsh introduces reusable tactile features and TacBench (Higuera et al., 2025); AnyTouch and AnyTouch 2 extend representation learning (Feng et al., 2025; 2026). Our force data are a downstream task. BADControl and the Phantom Force poster address different control and sensor attack surfaces (Burbano et al., 2026; Kong et al., 2026); our offline force estimates establish no controller or physical consequence.

Poisoning and curation. Naturally occurring subgroup attacks and their capacity dependence are established (Jagielski et al., 2020; Gupta et al., 2024). BadNets, BadEncoder and natural backdoors use distinct training-time access (Gu et al., 2019; Jia et al., 2022; Wenger et al., 2022); regression backdoors already motivate local targets and correctly labelled neighbors (Li et al., 2021). We study how contact grouping and verified-label reuse change the evaluation of these established poisoning and curation mechanisms.

BIG-LID studies nonlinear regression poisoning, SODA semantic classification backdoors, and DRMGuard model-level regression backdoor detection (Weerasinghe et al., 2021; Sun et al., 2024; Du et al., 2024). Their interfaces differ from our curation checks; we neither reproduce nor rank these methods. Our BayesClean I1 adaptations (Carnerero-Cano et al., 2026) share posteriors and compare supplied versus verified normalization (Appendices K–L); they do not reproduce the source benchmark or author code. Our clean fine-tuning control follows the recovery setting studied alongside pruning by Liu et al. (2018), without implementing Fine-Pruning. Numerical-label error detection (Zhou et al., 2023), structured-data validation and hierarchical inference (Roberts et al., 2017; Lee et al., 2023) motivate residual checks and trajectory-level units.

7 Scope and limitations

Our empirical conclusions concern label interventions in frozen-head pipelines on one selected public acquisition population. Indenter conditions were recorded separately, so geometry is confounded with condition-specific acquisition and force statistics. The two encoders extend representation coverage but share DINO-pretrained ViT families and differ in input interface and capacity. Repeated splits measure variation on these contacts, not across independently collected devices.

The curator trusts condition, batch and contact metadata, and corruption is confined to sharp candidates, enabling stratified calibration and category exclusion. Correct-row protection assumes exchangeable parent contacts; selected sources and acquisition shifts need separate assessment. Training-pool comparisons include induced scaling and sampling changes. The accepted-label and recovery studies test one construction and schedule; the I1-aware follow-up is a fixed 20-query heuristic, not a converged or worst-case attack. Metadata attacks, other conditions, independent devices and controllers remain untested.

8 Conclusion

Provenance constrains deletion cost but not label quality. Curation evaluations should report source allocation, correct-data loss and final utility under the same verified information.

AI USE STATEMENT

OpenAI Codex assisted with research questions and hypotheses, protocol design and critical review, data and implementation audits, mathematical statements and proofs, literature search, experimental software, execution and analysis, scientific figures, and manuscript drafting and revision. The assistance included identifying and correcting confounded controls before the formal comparison runs were executed. Numerical results are produced by the executed programs and retained prediction records, rather than estimated by a language model. Public tactile observations were not generated or visually modified by generative AI. Controlled target-label replacements are the explicitly specified experimental intervention. Mathematical derivations were checked separately and with exact finite-distribution calculations. The authors retain responsibility for the correctness, originality, interpretation, and final content of the submission.

ETHICS STATEMENT

This study examines a dual-use data-integrity problem using public tactile recordings and offline force estimators. It does not operate a robot, modify a deployed system, or demonstrate physical harm. The 5 N target denotes a model estimate, not a command or measured contact force. The experimental software makes label changes explicit and evaluates their detection and downstream effects. The data and pretrained weights remain subject to their upstream terms; acquisition scripts pin their provenance and do not redistribute the raw dataset or checkpoint. Any claim about physical consequence would require an independently evaluated controller and appropriate hardware limits.

REPRODUCIBILITY STATEMENT

The supplementary code specifies acquisition, preprocessing, frozen feature extraction, all group splits, sample and label assignments, calibration, training, and figure generation. Appendices give exact protocols and mathematical proofs. The execution artifact records release and file hashes, configurations, every seed's predictions and filtering decisions, and aggregate tables. Results can be recomputed from the saved predictions or regenerated from the public data. The studies use one selected two-batch dataset; repeated runs are not independent acquisitions. Software versions and measured CPU execution times are recorded with the artifact.

REFERENCES

AI at Meta. GelSight force estimation dataset card, 2024. URL https://huggingface.co/datasets/facebook/gelsight-force-estimation. Public TacBench dataset documentation; accessed September 13, 2026.

Luis Burbano, Hampei Sasahara, Ruoyu Song, Z. Berkay Celik, and Alvaro A. Cardenas. BAD-Control: Backdoor attacks against control systems. In *35th USENIX Security Symposium*, pp. 3379–3398. USENIX Association, 2026. URL https://www.usenix.org/conference/usenixsecurity26/presentation/burbano.

Javier Carnerero-Cano, Luis Muñoz González, Phillippa Spencer, and Emil C. Lupu. Stealthy poisoning attacks bypass defenses in regression settings. *arXiv preprint arXiv:2601.22308*, 2026. URL https://arxiv.org/abs/2601.22308.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Lingyu Du, Yupei Liu, Jinyuan Jia, and Guohao Lan. Defending deep regression models against backdoor attacks. *arXiv preprint arXiv:2411.04811*, 2024. URL https://arxiv.org/abs/2411.04811.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XToAemis1h.

Ruoxuan Feng, Yuxuan Zhou, Siyu Mei, Dongzhan Zhou, Pengwei Wang, Shaowei Cui, Bin Fang, Guocai Yao, and Di Hu. AnyTouch 2: General optical tactile representation learning for dynamic tactile perception. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.09617.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Isha Gupta, Hidde Lycklama, Emanuel Opel, Evan Rose, and Anwar Hithnawi. Fragile giants: Understanding the susceptibility of models to subpopulation attacks. *arXiv preprint arXiv:2410.08872*, 2024. URL https://arxiv.org/abs/2410.08872.

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 885–915, 2025. URL https://proceedings.mlr.press/v270/higuera25a.html.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732. URL https://projecteuclid.org/euclid.aoms/1177703732.

Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. Subpopulation data poisoning attacks. *arXiv preprint arXiv:2006.14026*, 2020. URL https://arxiv.org/abs/2006.14026.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE Symposium on Security and Privacy*, pp. 2043–2059, 2022. doi: 10.1109/SP46214.2022.9833644.

Zirui Kong, Youqian Zhang, and Sze Yiu Chau. POSTER: Phantom Force: Injecting adversarial tactile perceptions into embodied intelligence via EMI. In *ACM Asia Conference on Computer and Communications Security*, 2026. doi: 10.1145/3779208.3804894. URL https://arxiv.org/abs/2605.13492.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. doi: 10.1109/LRA.2020.2977257.

Yonghoon Lee, Rina Foygel Barber, and Rebecca Willett. Distribution-free inference with hierarchical data. *arXiv preprint arXiv:2306.06342*, 2023. URL https://arxiv.org/abs/2306.06342.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://www.stat.cmu.edu/~ryantibs/papers/conformal-jasa.pdf.

Xi Li, George Kesidis, David J. Miller, and Vladimir Lucic. Backdoor attack and defense for deep regression. *arXiv preprint arXiv:2109.02381*, 2021. URL https://arxiv.org/abs/2109.02381.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses*, pp. 273–294, 2018. URL https://arxiv.org/abs/1805.12185.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Bing Sun, Jun Sun, Wayne Koh, and Jie Shi. Neural network semantic backdoor detection and mitigation: A causality-based approach. In *33rd USENIX Security Symposium*, pp. 2883–2900. USENIX Association, 2024. URL https://www.usenix.org/conference/usenixsecurity24/presentation/sun-bing.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *arXiv preprint arXiv:1706.06991*, 2017. URL https://arxiv.org/abs/1706.06991.

Sandamal Weerasinghe, Tamas Abraham, Tansu Alpcan, Sarah M. Erfani, Christopher Leckie, and Benjamin I. P. Rubinstein. Closing the BIG-LID: An effective local intrinsic dimensionality defense for nonlinear regression poisoning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 3176–3184, 2021. doi: 10.24963/ijcai.2021/437.

Emily Wenger, Roma Bhattacharjee, Arjun Nitin Bhagoji, Josephine Passananti, Emilio Andere, Heather Zheng, and Ben Zhao. Finding naturally occurring physical backdoors in image datasets. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8af749935131cc8ea5dae4f6d8cdb304-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. GelSight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. doi: 10.3390/s17122762.

Hang Zhou, Jonas Mueller, Mayank Kumar, Jane-Ling Wang, and Jing Lei. Detecting errors in a numerical response via any regression model. *arXiv preprint arXiv:2305.16583*, 2023. URL https://arxiv.org/abs/2305.16583.

A Data Inventory and Provenance

The empirical study uses the selected six banks in Table 4. There are 101,072 raw image frames and 54,698 usable examples under the public loader's indexing convention before the deterministic per-trajectory cap. The retained 20,580 examples have legitimate normal-force labels between 0.0633026418 and 1.8024105021 N. Values in the table are rounded for display; mathematical support bounds are rounded outward.

Table 4: Independently verified selected-data inventory. $H \times W = 320 \times 240$ for the raw RGB images. Background indices refer to the condition/batch image bank.

Condition	Batch	Trajectories	Retained pairs	Maximum $\lvert F_z \rvert$ (N)	Background index
Flat	1	151	3,429	1.483017	0
Flat	2	147	3,372	1.169807	0
Sphere	1	148	3,342	1.802411	60
Sphere	2	154	3,473	1.461462	0
Sharp	1	153	3,545	1.331397	0
Sharp	2	151	3,419	1.765768	0
Total		904	20,580		

The dataset is pinned to revision `136db55485b6501605b1d2648ce0fe44740e302e` of `facebook/gelsight-force-estimation`. The acquisition program verifies all 29 compact image and metadata files against the release hashes. It excludes large duplicate `org_dataset` files. The independently checked inventory establishes that source-frame indices are not shared across contact trajectories within a bank, and that all six background references are outside every contact trajectory. A separate audit hashes all 29,599 distinct retained source images as both JPEG bytes and decoded RGB pixels. It finds no cross-trajectory duplicates of current images or ordered input pairs. Within-trajectory repetitions leave 19,370 unique current-frame contents and 20,242 unique ordered-pair contents; these repetitions remain inside their assigned trajectory partition. This prevents those specific forms of frame reuse; it does not make contacts independently acquired sensors or certify their physical timestamp alignment. We preserve the implemented public loader's convention, including its use of an offset-five length with current-frame index j.

The frozen encoder file is `dino_vitbase.safetensors` from `facebook/sparsh-dino-base`, revision `89e38eaade8eeedfa0f9c8857a76f57a2acd8206`. Its SHA-256 digest is

`169262848300790daeba9f380d04c4873ec37ca25b034bb8db35e1c28cb6f956`.

Necessary official model files are pinned to Sparsh source revision `fee6a05a97330eca014b59c71ac41596fe3974a7`, retained byte-for-byte with a file manifest and upstream license. Preprocessing was separately checked on all 65,536 uint8 pixel/background combinations, all 20,580 retained index/label tuples, and 60 real image pairs distributed across the six banks. The study uses float32 extraction; the supplementary precision comparison is a software diagnostic, not an alternative empirical feature bank.

B Experimental Protocols

The experiments are implemented in `run_attacks.py`, `run_defenses.py`, and `run_mitigation.py`. They use the independently extracted, pinned feature bank. These studies answer different questions and have separate partitions: conditional response to label corruption, detection of inconsistent supplied labels, and downstream performance after filtering and retraining. Seeds are $0, \ldots, 4$.

B.1 Common observations, features, and evaluation units

The data scope is the complete set of 904 recorded contact trajectories in the flat, sphere, and sharp banks of acquisition batches 1 and 2. Within each trajectory, at most 24 uniformly spaced

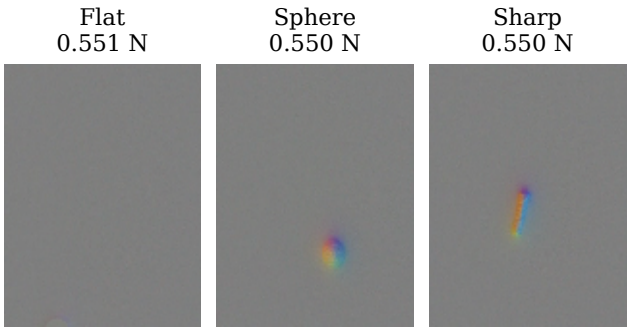

Figure 2: Actual current-frame difference images from the closest retained batch-1 sample to 0.55 N in each condition. Selection is deterministic and independent of model response. The flat contact lies near the lower image edge; no crop, contrast enhancement, or generative image modification is applied. These are illustrative observations, not matched geometric interventions. Exact image indices and forces are recorded with the figure script.

examples are retained. With upstream image indices I and force records F, the usable length is $m = \min\{|F|, |I| - 5\}$. Integer indices are obtained by uniformly spacing $\min\{m, 24\}$ points from 0 to $m - 1$. For retained index j, the current image is I_j, the earlier image is $I_{\max\{j-5,0\}}$, and the regression target is $|F_{j,z}|$ in newtons. Neither force labels nor predictions select the retained indices. Condition, batch, trajectory, retained index, and current/earlier source-frame indices are preserved in the feature bank. Condition, acquisition-batch, and trajectory identifiers are treated as verified metadata outside the evaluated attacker's control. The intervention changes selected response labels; it does not falsify source identities or condition assignments.

Preprocessing follows the public force-loader convention rather than a continuous-valued approximation. For RGB image I and the first recorded noncontact background B of its condition/batch bank, the difference tensor is

$$d(I, B) = \frac{1}{255} \left\lfloor 255 \,\mathrm{clip}\left(\frac{I - B}{255} + \frac{1}{2}, 0, 1 \right) \right\rfloor . \tag{6}$$

The floor represents conversion back to unsigned 8-bit pixels before tensor conversion; exact zero difference therefore becomes $127/255$. Landscape images are rotated clockwise to portrait orientation. Resizing is only applied if needed to obtain height 320 and width 240, using bilinear interpolation with antialiasing. The current difference image precedes the earlier image in the six-channel tensor. There is no ImageNet normalization.

The released Sparsh DINO ViT-B encoder is loaded strictly, frozen, and evaluated in FP32 on CPU (Higuera et al., 2025). Mean pooling averages the 300 normalized patch tokens returned by the official forward method; the register token is excluded. The result is a 768-dimensional feature saved in FP32. Source manifests pin the dataset files, checkpoint, model implementation, preprocessing, and feature-bank digest. Background references are fixed acquisition-specific calibration images, not fitted from force labels. Their use does not establish calibration on a new acquisition session: all three experiments concern the stated two-batch collection.

For any per-frame quantity h_i, let $\langle h \rangle_{\mathcal{G}} = |\mathcal{G}|^{-1} \sum_{g \in \mathcal{G}} |g|^{-1} \sum_{i \in g} h_i$. MAE, target error, prediction mean, and activation rates use this trajectory-balanced average. For target a, the principal response measures are

$$\mathrm{TE}_a = \langle |\widehat{y} - a| \rangle, \quad \mathrm{TB}_a = \langle \mathbf{1}\{|\widehat{y} - a| \le 0.1\,\mathrm{N}\} \rangle, \quad A_4 = \langle \mathbf{1}\{\widehat{y} \ge 4\,\mathrm{N}\} \rangle. \tag{7}$$

The target-band baseline is necessary because a legitimate estimate can already be close to an admissible target. Ordinary-contact A_4 is the non-trigger false-activation rate; neither it nor target-band success measures all regression damage. Paired errors compare predictions on identical held-out observations, for example $\Delta\,\mathrm{MAE} = \langle |\widehat{y}_p - y| - |\widehat{y}_0 - y| \rangle$. R^2 pools frames, so its weighting differs from trajectory-balanced MAE.

B.2 MATCHED SOURCE BUDGETS AND LABEL-MIXTURE INTERVENTIONS

Within each condition/batch stratum of U trajectories, a seeded random permutation assigns $\lfloor 0.2U \rfloor$ trajectories to test, the same number to validation, and the remainder to fitting. Every frame from a trajectory has the same assignment. The ordinary task-training set contains all retained flat/sphere fitting frames. Sharp fitting frames supply the controlled additional examples. Sharp validation labels and all test predictions are excluded from training-budget selection.

Head and optimization. The head is a $768 \to 512 \to 256 \to 1$ MLP with GELU hidden activations. Each feature coordinate is standardized using only the actually observed training inputs of that arm; the standard deviation is floored at 10^{-6}. Training uses AdamW, learning rate 10^{-3}, weight decay 10^{-4}, and batches of 256 indices sampled with replacement. The default loss is Smooth-L1 with $\beta = 0.5$. Update counts refer to optimizer steps, not full passes over the training set.

For each seed, an ordinary-only clean head is assessed after 1,000, 2,000, and 4,000 updates. The checkpoint with minimum trajectory-balanced ordinary validation MAE selects the update count; smaller counts break exact ties. Every subsequent arm for that seed uses that count without further selection. The same initialization seed is reset immediately before each head is created. The minibatch random stream is independently seeded by a hash of the seed and its fixed role. Identical input arrays therefore have identical scaling, initial parameters, and sampled row indices. Different-sized input arrays share the sampling rule, not identical sampled observations.

Source-count intervention. The controlled source-count arms each contribute exactly $k = 48$ sharp frames, drawn from 2, 12, or 48 source trajectories: respectively 24, 4, or 1 frame per source. All three arms use the same eligible pool of sharp fitting trajectories having at least 24 retained frames. This common restriction prevents the concentrated arm alone from excluding shorter trajectories. Within each acquisition batch, a fixed seeded ranking of eligible sources is reused across source counts, producing nested source selections. Counts are divided equally across the two batches, so every arm contributes exactly 24 poison frames from each batch. A separate fixed ranking of frames within each selected source is reused across source counts. Source selection does not depend on labels, feature distances, or model response.

Two uniform-frame comparators sample without replacement from the full sharp fitting bank: one selects 48 frames and the other selects $k = \mathrm{round}(0.05n/0.95)$ for n ordinary training frames. The latter targets a final-set fraction $k/(n+k)$ of 5% up to rounding. The uniform comparators can include shorter sources; their eligibility pool is therefore reported separately from the matched source-count comparison. For every strategy, the same selected rows are evaluated with fixed targets $a = 1.5$ N and $a = 5$ N. Each strategy also has an all-correct-label control with precisely the same ordered input array. This comparison isolates label replacement from the effect of adding sharp inputs or changing normalization. The ordinary-only reference is retained as a separate no-sharp baseline; it must not be substituted for the matched-input reference when interpreting the effect of label corruption.

Each run records the actual frame fraction, number of poison-contributing sources, eligible-source count, and fraction of available sharp fitting sources touched. The source share uses all distinct sources represented in that arm's training array as its denominator. These are provenance quantities, not independent-sample counts or measured costs of acquiring physical contacts.

Fixed-input label-mixture intervention. An independently seeded uniform sample of 240 sharp fitting frames is appended to the ordinary fitting inputs. Its order remains fixed. For $r \in \{0, 0.2, 0.5, 0.8, 1\}$, the first $240r$ selected sharp labels are replaced by 5 N; the remaining selected sharp labels stay correct. Thus input count, input order, normalization, and the sharp-source population are unchanged. Each mixture is trained with both Smooth-L1 and squared-error loss. Its matched reference is the $r = 0$ model trained with the same loss. The scalar r describes the selected sharp bank; a realized finite assignment need not have that corruption proportion in every local feature neighborhood. The update budget and other settings are inherited from the ordinary-only Smooth-L1 validation procedure, rather than tuned separately for squared error. This is a controlled loss comparison under shared settings, not a comparison of each loss's best achievable performance.

B.3 TRUSTED-REFERENCE SCREENING OF SUPPLIED LABELS

The screening study independently assigns approximately 60/20/20% of whole trajectories to fitting, calibration, and query partitions within each condition/batch stratum. Here each 20% count is rounded to the nearest integer and the remainder is used for fitting. The reference regressor sees *verified true labels from all three conditions*, including sharp contacts. Counts of trusted fitting and calibration trajectories and frames are exported separately; this is an explicit additional source of information.

The reference predictor is Ridge regression on standardized features. Its penalty is selected from $\{0.1, 1, 10, 100\}$ using a nested 80/20 group split inside the reference fitting partition, minimizing trajectory-balanced validation MAE and breaking ties toward the smaller penalty. Standardization is fitted inside the corresponding training split. After selection, both scaler and Ridge model are refitted on the complete outer fitting partition. Ridge fitting weights retained frames equally. Calibration and query examples never participate in fitting or penalty selection. Residual screening is a standard numerical-response consistency baseline (Zhou et al., 2023), not a new regression estimator.

For supplied labels $\widetilde{y}_i$, the two source scores are $S_g^{\max} = \max_{i \in g} |\widetilde{y}_i - \widehat{f}_{\mathrm{ref}}(z_i)|$ and $S_g^{\mathrm{mean}} = |g|^{-1} \sum_{i \in g} |\widetilde{y}_i - \widehat{f}_{\mathrm{ref}}(z_i)|$. Each is calibrated separately within condition and acquisition batch. For n_c clean calibration-source scores and $\alpha = 0.05$, define

$$h = \lceil (n_c + 1)(1 - \alpha) \rceil, \qquad q = \begin{cases} S_{(h)}, & h \le n_c, \\ +\infty, & h > n_c. \end{cases} \tag{8}$$

A source is flagged only when $S_g > q$. The infinite threshold is required when there are too few calibration sources for a finite threshold at the chosen level; with $\alpha = 0.05$, at least 19 are required. This is ordinary split calibration applied to entire trajectory objects. Its false-positive interpretation assumes exchangeability of clean calibration and query trajectories within the stated stratum, conditional on the fitted reference. It does not require independent frames within a trajectory, and it does not provide a future-device guarantee (Lee et al., 2023).

Every held-out sharp query trajectory is evaluated with its true labels and with a predetermined label replacement: the middle retained sample under natural numeric sample-ID ordering receives 1.5 N or 5 N. An additional arm replaces all retained labels in the trajectory. Neither its position nor its target depends on reference predictions. These are paired counterfactual annotations of identical inputs, not additional independent observations. Reports distinguish benign-sharp rejection, ordinary-condition false positives, corrupted-sharp recall, newly flagged trajectories, and the paired detection-rate change. Conditional label AUROC pools the paired clean/corrupted sharp scores and treats ties with half credit.

The feature-only comparison fits a scaler, PCA with at most 128 components, and Ledoit–Wolf covariance using only trusted flat/sphere fitting features. Its score is the mean squared Mahalanobis distance of a trajectory's frames; the threshold is calibrated on complete flat/sphere calibration trajectories separately by batch. It therefore measures novelty relative to ordinary contacts. Replacing labels cannot change this score on the same inputs; corrupted-sharp recall and benign-sharp rejection are identical by construction. The accompanying label-range audit flags any source containing a supplied label above 3 N. Its reported AUROC uses the unthresholded maximum supplied label as the score. High ranking accuracy does not imply detection at the fixed 3 N threshold. That diagnostic range is not an independently certified sensor bound.

B.4 FILTERING FOLLOWED BY VICTIM RETRAINING

The end-to-end study uses a separate four-way split within every condition/batch stratum: 20% trusted reference fitting, 20% trusted calibration, 20% held-out testing, and the remaining approximately 40% for victim fitting. The three 20% counts are individually rounded. Reference Ridge selection and feature-novelty fitting follow the procedures above, using only the trusted partitions. The victim does not train on those reference trajectories. The already fitted trusted Ridge is also evaluated directly on the same held-out ordinary, sharp, and combined test sets. Its trajectory-balanced MAE and pooled R^2 provide a reference-only context baseline: they show the predictive value of the verified data available to the screening procedure. This evaluation uses cached predictions, requires no additional model fitting, and is reported separately from trained victim arms.

For n ordinary victim-fitting frames, set $k = \mathrm{round}(0.05n/(1 - 2 \cdot 0.05))$. Select $2k$ sharp victim-fitting frames uniformly without replacement; label the first k with target $a \in \{1.5, 5\}$ N and leave the other k correctly labelled. The actual poison fraction is $k/(n + 2k)$. Poison and cover observations are disjoint, while their source trajectories may overlap; both are recorded. The original ordered candidate input bank is identical across targets and filtering methods. Its clean baseline uses the same $n + 2k$ input rows with every force label correct, including all selected sharp examples.

Five policies process the candidate bank before training: retain every row; remove every sharp row using the verified condition metadata; remove only rows with supplied labels above 3 N; remove all candidate rows from sources exceeding the feature-novelty threshold; or remove all candidate rows from sources exceeding the maximum-residual threshold. Source scores are computed from the rows actually contributed to the candidate bank. Condition exclusion is the transparent reference for rejecting the entire known category; its effect on legitimate sharp prediction is measured. Consequently a flagged source can lose correctly labelled cover rows as well as poison rows. For every rule, the artifact also evaluates filtering decisions on the identical candidate inputs with all labels restored to their true values. In the original mitigation study, this paired audit measures collateral removal without changing candidate-source composition and does not fit an additional filtered-clean head. The adaptive follow-up in Appendix F.3 adds a distinct retained-input correct-label control.

The calibration objects contain all retained frames of trusted sources, whereas a candidate sharp source can contribute only a subset. Mean feature scores therefore need not have the same distribution. A clean subset's maximum residual cannot exceed its full-source maximum, but uniform frame sampling also makes the probability of selecting a source depend on its retained size. For these reasons, Equation 8 supplies the specified operating threshold here; we do not claim an unconditional 5% rejection guarantee for the selected candidate bank. The paired all-correct-label audit reports its actual collateral rejection under the implemented sampling rule.

After filtering, each victim head and the clean same-input-bank baseline are trained for exactly 2,000 optimizer updates using the shared MLP, Smooth-L1, and optimization settings. This budget is fixed independently of predictions. Every call resets the same seed immediately before head construction, giving identical initial parameter tensors across that seed's arms. Scalers are fitted to each arm's retained inputs; therefore equal initialization does not imply an equal raw-input predictor when filtering changes scaling. The artifact saves the shared initial state and its digest, each trained checkpoint and scaler, the input indices and supplied labels, row masks, source flags, and test predictions.

Filtering quality is assessed by poison-row removal recall, poison-source flag recall, ordinary-row and benign-cover removal, entirely removed sources, and remaining contamination. Prediction metrics are then computed on untouched test trajectories, separately for ordinary contacts, sharp contacts, and all legitimate conditions. Paired MAE and target-band changes use the clean same-input-bank baseline. These results jointly characterize detection, collateral data loss, and downstream utility; a screening score alone is not treated as successful mitigation.

B.5 Scope, reproducibility, and interpretation

The methods are established curation baselines under explicit reference-data assumptions. In particular, the residual method knows verified sharp force labels. Feature novelty uses only ordinary reference features, and the range audit assumes an admissible-label bound. These differences are part of the comparison, not hidden advantages. DRMGuard instead examines an already trained regression model and reverse-engineers a trigger using a small clean reference set (Du et al., 2024). Our pre-adaptation data curation experiment neither implements DRMGuard nor establishes comparative superiority over it.

The new studies export individual seeded records and descriptive means and sample standard deviations. Seeds change partitions, selected input rows, initialization, and optimization streams within one reused finite collection; they are not independently acquired sensors, gels, or laboratories. No significance claim follows from the run count alone. Source manifests, configuration records, selected indices, supplied labels, prediction arrays, and checkpoints make the reported comparisons inspectable. The source-budget experiment also records feature, implementation, and configuration digests before allowing a resumed run. These protocols specify what is tested; conclusions must be restricted to the independently executed results.

B.6 WHOLE-BATCH TRANSFER WITH AN INPUT-MATCHED REFERENCE

The batch-shift stress test is implemented independently in `run_batch_shift.py`. We train on acquisition batch 1 and evaluate on the entirety of batch 2, then reverse the roles. Within the source batch, flat and sphere trajectories are separately permuted; the first $\lfloor 0.2U \rfloor$ of each condition's U trajectories form clean validation and the remainder form ordinary training data. All source-batch sharp trajectories are eligible for selection. The target batch supplies its own noncontact background image, as required by the upstream loader; this unlabelled sensor calibration is allowed. No target-contact observation is used to select examples, fit the training feature mean or variance, choose a head, or select the update budget.

An ordinary-only head selects 1,000, 2,000, or 4,000 optimizer updates using source-batch clean validation MAE, with smaller counts breaking ties. For n ordinary training frames, a fixed seeded draw selects $k = \mathrm{round}(0.05n/0.95)$ sharp frames uniformly without replacement from the source pool. These selected frames produce three heads with exactly the same ordered inputs: one retains all true labels, one replaces the selected sharp labels by 1.5 N, and one replaces them by 5 N. The feature mean and standard deviation, initialization seed, minibatch-index random stream, optimizer, loss, and update count are identical across these three heads. Only the supplied labels change. The correctly labelled head is therefore the reference for attributing prediction changes to label corruption; the ordinary-only head separately measures adaptation without sharp coverage.

On the entire target batch, we report trajectory-balanced ordinary and sharp MAE, target error and target-band response, ordinary false activation above 4 N, and pooled ordinary R^2. For each poisoning target, its input-matched reference supplies the target-band baseline and the paired changes in ordinary and sharp MAE and prediction mean. We additionally retain comparisons against the ordinary-only head. The actual frame fraction, number of selected sharp source trajectories, eligible-source count, and source share are recorded for each run. Training, validation, and test trajectory identities, selected frame indices, supplied labels, predictions, and checkpoints are saved.

The five seeds $0, \ldots, 4$ vary the source-batch clean split, sharp draw, initialization, and optimization. Means and sample standard deviations summarize this variation within each fixed direction. They do not turn two acquisition batches into five independent sessions, identify a cause of acquisition shift, or establish transfer across sensors or laboratories.

C ANALYTICAL REFERENCE CALCULATIONS

C.1 FINITE-POPULATION SOURCE ACCOUNTING

If sharp training trajectories retain $m_1, \ldots, m_{U_T}$ frames and $M = \sum_j m_j$, uniform sampling of k frames without replacement gives

$$\mathbb{E}[U_P] = \sum_{j=1}^{U_T} \left(1 - \frac{\binom{M-m_j}{k}}{\binom{M}{k}} \right). \tag{9}$$

The capacity bound is $U_P \geq \lceil k / \max_j m_j \rceil$. These quantities describe represented source counts; neither predicts an attack response. The fixed-count and uniform-sampler protocols are specified separately in Appendix B.2.

Let I_j indicate that at least one of the m_j retained frames from source j is sampled in a uniform size-k subset of an M-frame bank. There are $\binom{M-m_j}{k}$ subsets avoiding source j and $\binom{M}{k}$ total subsets. Therefore $\mathbb{E}I_j = 1 - \binom{M-m_j}{k} / \binom{M}{k}$, with the numerator zero when $M - m_j < k$. Summing $U_P = \sum_j I_j$ proves Equation 9 without assuming independent source indicators. Capacity alone implies $k \leq U_P \max_j m_j$, giving the stated lower bound. Our source-count experiment tests the outcome that this counting argument cannot predict. Exact expectations are evaluated separately using each run's source sizes and integer poison count; averages are not substituted into the combinatorial formula.

C.2 THE FEATURE-ONLY RESTRICTION

For probability laws Q_1, Q_0 and any measurable $r \in [0,1]$, the layer-cake identity gives

$$\mathbb{E}_{Q_1} r - \mathbb{E}_{Q_0} r = \int_0^1 [Q_1\{r > t\} - Q_0\{r > t\}]\, dt. \tag{10}$$

The absolute value is bounded by the supremum of the probability difference over measurable sets, namely total variation. A randomized rule is handled by its conditional rejection probability. If uniformly assigned corruption is independent of features, the conditional laws coincide in the assignment distribution. Finite realized subsets can still differ. Our paired-label diagnostic is stronger for its fixed query inputs: every feature score is identical before and after label replacement, so the two empirical score multisets agree and the label-discrimination AUROC is exactly $1/2$ with half credit for ties. A score fitted using evaluated corruption assignments would not meet the fixed-rule premise.

C.3 CONDITIONAL LABEL COMPOSITION UNDER SMOOTH-L1

This calculation explains why benign trigger coverage is a substantive intervention. It is an application of classical Huber contamination and clipped-score estimation (Huber, 1964; Sun et al., 2017), not a new robustness theorem for neural networks. Unlike a two-point-label illustration, the first statement permits an arbitrary legitimate conditional force distribution with bounded support.

Fix a representation z in the trigger condition. Let the legitimate force $Y \mid Z = z$ have distribution P_z supported on $[\ell, u]$, and let the observed force have distribution $(1-\pi)P_z + \pi\delta_a$. Here π is the *conditional* corrupted-label fraction, not the fraction of all training frames. With $\beta > 0$, write the Smooth-L1 loss and score as

$$\rho_\beta(e) = \begin{cases} e^2/(2\beta), & |e| \le \beta, \\ |e| - \beta/2, & |e| > \beta, \end{cases} \qquad \psi_\beta(e) = \rho'_\beta(e) = \mathrm{clip}(e/\beta, -1, 1). \tag{11}$$

This is Huber loss divided by β, so the two losses have identical minimizers for fixed β. The conditional population risk is

$$R_z(t) = (1-\pi)\mathbb{E}_{P_z}\rho_\beta(t-Y) + \pi\rho_\beta(t-a). \tag{12}$$

Proposition 2 (Bounded-label contamination bounds)**.** *The minimizer set $\mathcal{M}_z = \arg\min_{t\in\mathbb{R}} R_z(t)$ is nonempty. For every $a \in \mathbb{R}$ and $0 \le \pi < 1/2$, every $t^* \in \mathcal{M}_z$ satisfies*

$$\ell - \frac{\beta\pi}{1-\pi} \;\le\; t^* \;\le\; u + \frac{\beta\pi}{1-\pi}. \tag{13}$$

For $1/2 < \pi \le 1$ and every $a \in \mathbb{R}$,

$$a - \frac{\beta(1-\pi)}{\pi} \;\le\; t^* \;\le\; a + \frac{\beta(1-\pi)}{\pi}. \tag{14}$$

In particular, when $a > u$, the bounds refine to $t^* \ge \ell$ in (13) and $t^* \le a$ in (14). No target-separation assumption is needed for these bounds.

Proof. The risk is finite, continuously differentiable, convex, and coercive: all labels are bounded and ρ_β grows linearly at infinity. It therefore attains its minimum. Boundedness of ψ_β permits differentiation under the expectation, giving

$$R'_z(t) = (1-\pi)\mathbb{E}_{P_z}\psi_\beta(t-Y) + \pi\psi_\beta(t-a). \tag{15}$$

For $\pi < 1/2$, put $q = \pi/(1-\pi) < 1$. If $t > u + \beta q$, then for all $Y \le u$, $\psi_\beta(t-Y) \ge \psi_\beta(t-u) > q$. Consequently $R'_z(t) > (1-\pi)q - \pi = 0$, excluding a minimizer. The same argument below $\ell - \beta q$, with all signs reversed, excludes a minimizer there. For $\pi > 1/2$, put $q = (1-\pi)/\pi < 1$. If $t < a - \beta q$, then $\psi_\beta(t-a) < -q$, whereas the clean score is at most 1, so $R'_z(t) < (1-\pi) - \pi q = 0$. The symmetric argument excludes $t > a + \beta q$. Finally, if $a > u$, all positive-weight component labels belong to $[\ell, a]$; the derivative is strictly negative below ℓ and strictly positive above a. □

Table 5: Analytical scalar-risk consequences, not sensor experiments. Here $v = \operatorname{ess\,sup}(Y \mid Z = z) \leq 1.802411$ N. The last column concerns the population minimizer only.

Conditional poison π	Minimizer constraint (N)	$t^* \geq 4$ N?
0	$t^* \leq 1.802411$	Impossible
0.2	$t^* \leq 1.927411$	Impossible
0.5	$\mathcal{M}_z = [v + 0.5, 4.5]$	Depends on minimizer selection
0.8	$t^* = 4.875$	Certain
1	$t^* = 5$	Certain

Sharpness. The minority upper bound is attained by $P_z = \delta_u$ whenever $a - u \geq \beta/(1-\pi)$; the symmetric lower bound is also attainable. More generally, for $\pi > 1/2$ and $a - u \geq \beta/\pi$, the minimizer is *exactly*

$$t^* = a - \frac{\beta(1-\pi)}{\pi}. \tag{16}$$

At this value every legitimate residual is at least β, while the poison residual lies in $[-\beta, 0]$, and the score is zero. The positive-weight poison component is strictly convex locally at this point, which makes this global minimizer unique.

Proposition 3 (Equal mixture and complete two-point solution)**.** *Let $v = \operatorname{ess\,sup}_{P_z} Y$. If $\pi = 1/2$ and $a - v > 2\beta$, then*

$$\mathcal{M}_z = [v + \beta, a - \beta]. \tag{17}$$

If $P_z = \delta_{y_0}$ with $d = a - y_0 > 0$, the complete solution is

$$\mathcal{M}_z = \begin{cases} \{y_0 + \pi d\}, & d \leq \beta/\max\{\pi, 1-\pi\}, \\ \{y_0 + \beta\pi/(1-\pi)\}, & \pi < 1/2,\ d > \beta/(1-\pi), \\ [y_0 + \beta, a - \beta], & \pi = 1/2,\ d > 2\beta, \\ \{a - \beta(1-\pi)/\pi\}, & \pi > 1/2,\ d > \beta/\pi. \end{cases} \tag{18}$$

The cases include $\pi = 0$ and $\pi = 1$ without division by zero.

Proof. At equal mixing, every $t \in [v + \beta, a - \beta]$ gives clean score 1 almost surely and poison score -1, hence zero derivative. For $t < v + \beta$, the definition of essential supremum gives positive probability of $Y > t - \beta$, so the expected clean score is strictly less than 1. Because $t < a - \beta$, the poison score is -1, making the derivative strictly negative. For $t > a - \beta$, all clean scores are 1 and the poison score is strictly greater than -1, making the derivative strictly positive. This proves the exact plateau. For a point mass, every minimizer lies between y_0 and a. If both residuals are in their quadratic regions, the score equation gives $t = y_0 + \pi d$; feasibility is precisely $d \max\{\pi, 1-\pi\} \leq \beta$. Outside that regime and with $\pi < 1/2$, the target residual is saturated and the legitimate residual is quadratic, giving $(1-\pi)(t - y_0)/\beta - \pi = 0$. For $\pi > 1/2$ the roles reverse. Equal mixing is the plateau case already established, completing all regimes. □

Analytical values for the fixed-input label-mixture study. For $a = 5$ N, $\beta = 0.5$ N, and the descriptive retained-data upper bound $u = 1.802411$ N (rounding the retained maximum $1.802410502\ldots$ N upward), Table 5 follows directly from the propositions. These are predictions of the *conditional population oracle*; they are not measured neural-network outputs or measured ASRs. The bound u is an empirical support description and is not a certified operating range. Since $a - u > 2\beta$, every conditional distribution obeying this bound has a nontrivial equal-mixture plateau.

Why the fixed input bank matters. Suppose the selected sharp bank has M fixed feature–true-label pairs (z_i, y_i) and a uniformly selected size-k subset receives label a. For every fixed predictor g, expectation over this assignment gives

$$\mathbb{E}_A \widehat{R}_A(g) = \frac{1}{M} \sum_{i=1}^{M} \left[(1 - k/M)\rho_\beta(g(z_i) - y_i) + (k/M)\rho_\beta(g(z_i) - a) \right]. \tag{19}$$

Each index has the same marginal poisoning probability k/M; independence between indices is unnecessary. The corresponding population assumption is $A \perp (Z, Y) \mid T$, where T denotes the selected trigger condition. Then $\pi = \Pr(A = 1 \mid Z = z, T)$ is constant within that condition, while the legitimate force law may still vary with z. If the assignment depends on features but obeys $A \perp Y \mid (Z, T)$, π must instead be replaced by $\pi(z)$. If selection also depends on the true force given z, the uncorrupted component need not retain P_z and must be specified separately. A small overall frame poison rate does not bound local label corruption.

The derivation minimizes conditional population risk, or expected empirical risk for independently chosen scalar outputs. It does not interchange argmin and expectation over label assignments. A trained MLP shares parameters across inputs and also has finite data, regularization, and optimization error. In particular, one fixed assignment can create locally unbalanced corruption, and a flexible head may fit it. At $\pi = 1/2$ the scalar objective itself does not choose an output along the plateau. Therefore the table motivates controlled label-composition experiments; it does not certify robustness, forecast their exact ASR, or attribute all observed behavior to the loss alone.

Computational verification. The accompanying `verify_loss_theory.py` enumerates the exact zero set of the clipped-score derivative using rational arithmetic for finite legitimate distributions. It checks both endpoints of every minimizer interval, separation boundaries, unequal mixture weights, arbitrary target locations, equal-mixture plateaus, and the complete point-mass formula. Its numerical checks are a consistency audit of the derivation, not additional tactile learning experiments.

D EXACT COLLATERAL FRONTIER FOR CONTACT DELETION

Scope. The inputs to this analysis are the candidate-row records already used in the mitigation experiment. Within source j, p_j counts corrupted-label rows and b_j counts correctly labelled sharp rows. A source can contain both. The source partition is fixed, and whole-contact deletion removes all candidate rows from a selected source. The objective counts sharp cover; ordinary false rejections are separate. The oracle sees actual poison membership, which is unavailable to a practical detector. It provides a comparison for deletion cost rather than achievable detection accuracy or post-training regression error.

Uniform-sampler expectation. If k poison rows and $c \geq 1$ cover rows are sampled uniformly and disjointly from M sharp rows in contacts of sizes m_j, with $k + c \leq M$, the expected complete-recall collateral is

$$\mathbb{E}[L_{\text{contact}}] = \sum_j \frac{m_j}{M} \left(1 - \frac{\binom{M-m_j}{k}}{\binom{M-1}{k}} \right). \tag{20}$$

Proof of Proposition 1. Every source with $p_j > 0$ must be deleted to remove all K poisoned rows. Those choices remove $\sum_{j:p_j>0} b_j$ cover rows. Deleting exactly these sources attains the lower bound, since deleting a poison-free source cannot reduce the loss. For the sampling identity, select one of the c cover rows uniformly. Its marginal distribution is uniform among all M pool rows, so it belongs to source j with probability m_j/M. Conditional on this cover-row identity, the k poison rows are a uniformly random k-subset of the remaining $M - 1$ rows. The event of no poison from source j has probability $\binom{M-m_j}{k}/\binom{M-1}{k}$. Its complement is exactly the event that this cover row is forced out by a complete-recall whole-contact rule. Averaging across source membership and then across the c exchangeable cover rows proves Equation 20. This argument does not assume independent cover indicators. The convention $\binom{n}{k} = 0$ for $n < k$ handles contact exhaustion. The condition $c \geq 1$ ensures $k \leq M - 1$ and a nonzero denominator. □

Deterministic frontier. For every integer $t \in \{0, \ldots, K\}$, define

$$C(t) = \min_{d \in \{0,1\}^J} \left\{ \sum_j b_j d_j : \sum_j p_j d_j \geq t \right\}. \tag{21}$$

This is the exact minimum cover-row loss subject to removing at least t poison rows. Every permitted decision is one feasible binary vector, and a minimizer selects a valid subset of source contacts.

Poison-free sources can be omitted. For the remaining contacts, let $D_i(s)$ be the smallest cover loss that removes exactly s poison rows among the first i contacts. Initialize $D_0(0) = 0$ and all other states to infinity, then apply

$$D_i(s) = \min\{D_{i-1}(s),\ b_i + D_{i-1}(s - p_i)\}, \qquad C(t) = \min_{s \geq t} D_J(s), \tag{22}$$

where negative-index states are infeasible. The standard dynamic program takes $O(JK)$ time. We retain an attaining subset, verify its poison and cover sums independently, and export witnesses at requested recalls of $0, 25, 50, 75, 90, 95$, and 100%. The integer required count is the ceiling of requested recall times K. The full per-seed curve contains all integer counts; it is not inferred from the observed detector thresholds.

Fractional relaxation. Allowing $d_j \in [0, 1]$ gives a lower bound on Equation 21. Sort positive-poison contacts by b_j/p_j and take contacts in this order, possibly using a fraction of the boundary contact. The usual exchange argument proves optimality: transferring equal poison-removal mass from a higher-ratio selected contact to a lower-ratio unfilled one never increases cost. Here a fraction represents a probability of deleting an entire contact, not permission to remove arbitrary rows from it. The relaxed solution is attainable under constraints on *expected* recall and expected cover loss; it need not attain the requested recall in each realization. We use the integer frontier as the primary benchmark and export the fractional values separately. At recall one the two endpoints coincide.

Recorded-bank result and its limits. Across five seeds, $K = B$ averages 304.0 ± 1.87 rows from a sharp victim-fit pool of 2753.8 ± 24.87 rows and exactly 121 sources. Poison occurs in 111.6 ± 2.07 sources. The complete-recall lower bound is 284.8 ± 7.09 cover rows, or $93.6873 \pm 2.4142\%$. The residual-maximum contact filter removes exactly these cover rows in every seed. This does not say it is optimal for ordinary-data retention, nor that a lower-loss filter has better retrained regression error. The uniform-sampler expected bound is 92.2009%; its 0.1285 percentage-point standard deviation is variation of the computed expectation across the five source inventories, not a confidence interval or the sampling variation of realized collateral.

Independent computational checks. The audit reads candidate-row CSV files and source inventories directly; it does not import original detector or regression-metric functions and fits no model. It verifies 1,525 empirical integer-frontier points and all 50 original filter decisions, checking groupwise decisions against the relevant oracle constraint. The 5 N range audit deletes individual rows and is explicitly outside the whole-contact comparison. A separate exhaustive subset enumeration checks 1,559 toy frontier values and their fractional lower bounds. An exhaustive finite-urn calculation verifies 58 combinations of pool partition, poison count, and disjoint cover count against Equation 20. The script exports per-source inventories, full frontiers, attaining subsets, detector operating points, and SHA-256 hashes of every input CSV.

D.1 Row deletion under contact-level calibration

Fix the trusted fitting data and hence h, and fix a condition/batch stratum. Suppose the n_{cal} calibration trajectories and one new entirely correct trajectory are exchangeable as full trajectories. Their maximum residual scores are then exchangeable even if frames within a trajectory are arbitrarily dependent. Write $k = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$. If $k > n_{\text{cal}}$, the $+\infty$ threshold rejects no finite score. Otherwise, random tie-breaking makes the rank of the new score uniform on $\{1, \ldots, n_{\text{cal}} + 1\}$; using strict exceedance without random tie-breaking is conservative. Thus

$$\Pr\{E_{\text{new}} > E_{(k)}\} \leq \frac{n_{\text{cal}} + 1 - k}{n_{\text{cal}} + 1} \leq \alpha. \tag{23}$$

For the row rule, rejection of any row in the contact is equivalent to its maximum residual exceeding the threshold, proving the stated guarantee. Any subset of the same contact has a no-larger maximum, so row selection within that contact does not increase this false-rejection event.

Corollary: correct-row protection under label-only corruption. Let g denote the original, entirely correct contact in the preceding exchangeability assumption, before any labels are altered. A contributor may replace any subset of its labels while leaving its inputs, source membership and

metadata unchanged. Let $C_g \subseteq g$ contain the rows whose supplied labels remain equal to their true labels. The contributor may choose this subset and the replacement labels using the contact, the reference and the calibration records. The reference h and the threshold q are held fixed during this intervention; the probability below still averages over the calibration sample and the original contact, conditional only on the separate reference fit and the specified stratum. Then

$$\Pr\{D_{\text{row}} \cap C_g \neq \varnothing\} \leq \alpha. \tag{24}$$

Indeed, for each realization, correctly labelled rows have unchanged residuals, so

$$\{\exists i \in C_g : |Y_i - h(Z_i)| > q\} \subseteq \{\max_{i \in g} |Y_i - h(Z_i)| > q\}. \tag{25}$$

The rank bound on the right proves the claim, even for an adaptively selected C_g. This is a direct corollary of standard calibration and set inclusion, not a new conformal method. Whole-contact deletion lacks the same guarantee under corruption: one altered label can flag the contact and discard every correct row, even when all original residuals are below q.

The corollary does not condition on a contact having been attacked, on the realized calibration threshold, or on membership in a selected candidate bank. Selecting which parent contacts are supplied, including selection induced by frame sampling, can destroy the required exchangeability. Acquisition shift can do so as well. Consequently the mixed-contact candidate bank in this study still requires the reported empirical collateral-loss measurements; the corollary does not certify those observed rates.

Exact synthetic check. An independent rational-arithmetic enumeration checks 48 finite-population settings, including tied scores and infinite thresholds, and 77,120 supplied-label assignments. It verifies the deterministic inclusion above and the resulting marginal bound without using experimental predictions. A simple separating example has 20 exchangeable contacts with three true labels $(r/100, r/200, r/300)$, $r = 1, \ldots, 20$, and $h = 0$. Randomly hold out one contact, calibrate at $\alpha = 0.05$ on the other 19, and replace only its second label by $q + 1$. Whole-contact deletion then discards correct rows with probability one; row deletion does so only for $r = 20$, with probability $1/20$. This illustrates the corollary's distinction, not an empirical attack result. The script and exact outputs accompany the source package.

Equation 4 immediately implies set inclusion of deleted rows. It is not a dominance theorem for all outcomes: whole-contact deletion can remove otherwise undetectable poisons sharing a source with a detectable one; retaining more correctly labelled examples need not improve a particular finite training run. We consequently retrain both policies and record both poison recall and downstream errors. The rule is a controlled use of an existing residual score and standard split calibration, not a new conformal inference method.

E Complete Executed Results

Every table reports the mean and sample standard deviation of all five seeds. No test response is used to select a displayed run. The tables refer to the independently executed protocols above; none combines results from the previous manuscript or the incomplete timing pilot. Rates use percentages unless stated otherwise. Positive paired MAE change denotes degradation.

E.1 Primary matched-input controls and source budgets

Table 6 gives absolute errors for the ordinary-only head, each all-correct source-sampling control, and both label targets. Table 7 supplies target error, matched-reference target-band rates, and changes relative to the same-input correct-label head. The zero-sharp ordinary-only baseline is a different training distribution; its error is not the reference for the matched label-change estimate. Table 8 reports realized source quantities.

Source-allocation outcomes and sampler references. The ordinary-only head reaches $0.0277 \pm 0.0034\,\text{N}$ ordinary MAE and pooled $R^2 = 0.947 \pm 0.017$, with no observed sharp or ordinary $\geq 4\,\text{N}$ activation. Our controlled arms each contribute exactly 48 poisoned frames, approximately 0.58%

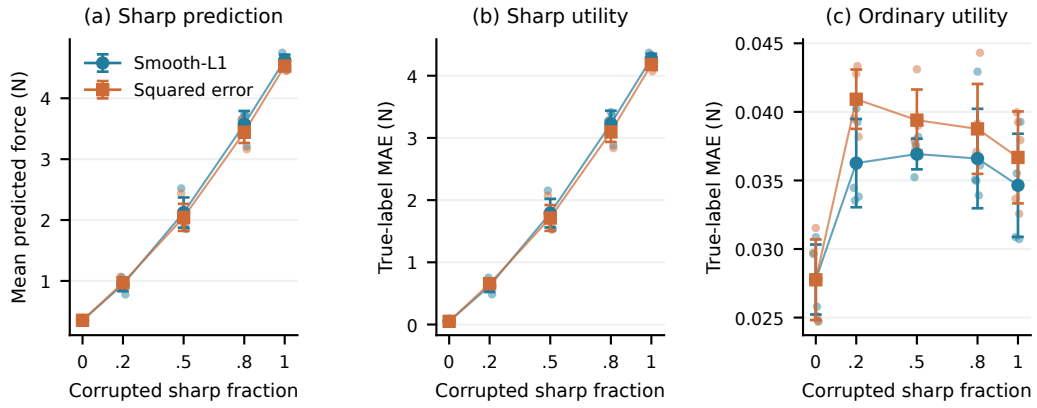

Figure 3: Label composition on a fixed 240-frame sharp bank. Small points show all five runs; markers and bars are means and sample SD. Lines connect the tested mixtures. Ordinary inputs remain fixed and each loss has its own correct-label reference at $r = 0$. The scalar Huber reference is an idealized risk calculation, not a fitted curve or guarantee for these neural heads.

of the final fitting set. Their source coverage ranges from 2/184 (1.1%) to 48/184 (26.1%) available sharp trajectories.

For the 5 N target, mean sharp exceedance rises from $5.3 \pm 6.0\%$ with 2 sources to $32.3 \pm 8.9\%$ with 12 and $45.9 \pm 12.6\%$ with 48 (Figure 1a). The allocation across contacts therefore matters beyond the number of poisoned frames. The uniform-48 reference represents 42.8 ± 2.9 sources and yields $54.0 \pm 5.9\%$ exceedance. Its pool includes shorter trajectories, so it is a separate sampler reference. Equation 9 predicts 42.45 represented sources, not a response.

Changing labels to 1.5 N also degrades true force estimation without crossing 4 N. Relative to the matched correct-label heads, sharp MAE increases by 0.119 ± 0.068, 0.435 ± 0.119, and 0.657 ± 0.039 N for 2, 12, and 48 sources. At 48 sources the target-band rate is $16.5 \pm 1.7\%$, versus zero for its correct-label control, while ordinary MAE increases by 0.0027 ± 0.0034 N. This is a condition-specific regression failure that the out-of-range activation metric misses. The 1.5 N construction passes the fixed range audit; this fact establishes range membership, not conditional plausibility or general curation stealth. Tables 6–8 include every matched control, target error, and the uniform 5% reference.

Second encoder. ImageNet DINO ViT-S/16 on the same current contact frames shows a larger response even with two sources: 5 N exceedance is 59.6% at 2 sources and 87.9% at 48; paired sharp-MAE increase at 1.5 N is 0.723 and 1.003 N, respectively (Figure 1). In its separate curation study, contact versus row residual deletion loses 93.69% versus 0.33% of sharp cover, with full poison recall for both; sharp MAE is 0.074 versus 0.037 N. The deletion-cost pattern recurs, but two sources already produce a strong response: broad source access is not necessary in every encoder setting. Appendix G.3 gives all controls and the changed single-frame, CLS-token and head interfaces.

E.2 Label composition matters at fixed inputs

The fixed 240-frame sharp bank changes only corruption fraction and loss. Under Smooth-L1, mean sharp prediction rises from 0.352 ± 0.022 N at $r = 0$ to 0.942 ± 0.112 at $r = .2$ and 4.627 ± 0.091 N at $r = 1$. MSE shows a similar empirical dependence at the shared settings. The finite head does not follow the ideal scalar solution: at $r = .8$ its mean is 3.573 N, versus the conditional-risk reference of 4.875 N. Some held-out predictions exceed 4 N even at $r = .2$. Figure 3, Table 9, and Appendix E.3 report complete curves and training errors. They establish a label-composition effect, not an automatic neural robustness guarantee from Smooth-L1. Table 9 compares both losses on the same 240 sharp inputs. The corruption fraction refers to this sharp bank, while ordinary

Table 6: Matched-input attack study. Values are means ± sample SD across five seeds. MAE is in newtons. Ordinary R^2 pools frames; MAEs balance trajectories. A_4 is the sharp-contact ≥ 4 N response rate. Correct controls share the exact inputs of their corresponding target arms.

Sampling	Labels	Ord. MAE	Sharp MAE	Ord. R^2	A_4 (%)
Ordinary only	Correct	0.028 ± 0.003	0.097 ± 0.018	0.947 ± 0.017	0.0 ± 0.0
2 sources	Correct	0.028 ± 0.004	0.094 ± 0.017	0.943 ± 0.020	0.0 ± 0.0
2 sources	1.5 N	0.029 ± 0.003	0.213 ± 0.071	0.944 ± 0.017	0.0 ± 0.0
2 sources	5 N	0.032 ± 0.005	0.925 ± 0.357	0.913 ± 0.048	5.3 ± 6.0
12 sources	Correct	0.028 ± 0.004	0.081 ± 0.014	0.943 ± 0.019	0.0 ± 0.0
12 sources	1.5 N	0.030 ± 0.002	0.516 ± 0.122	0.937 ± 0.013	0.0 ± 0.0
12 sources	5 N	0.033 ± 0.006	2.557 ± 0.457	0.877 ± 0.066	32.3 ± 8.9
48 sources	Correct	0.028 ± 0.004	0.067 ± 0.006	0.944 ± 0.020	0.0 ± 0.0
48 sources	1.5 N	0.031 ± 0.002	0.724 ± 0.040	0.934 ± 0.011	0.0 ± 0.0
48 sources	5 N	0.034 ± 0.005	3.214 ± 0.417	0.822 ± 0.164	45.9 ± 12.6
Uniform 48	Correct	0.028 ± 0.003	0.068 ± 0.010	0.945 ± 0.018	0.0 ± 0.0
Uniform 48	1.5 N	0.031 ± 0.003	0.767 ± 0.058	0.934 ± 0.009	0.0 ± 0.0
Uniform 48	5 N	0.036 ± 0.002	3.424 ± 0.280	0.806 ± 0.091	54.0 ± 5.9
Uniform 5%	Correct	0.027 ± 0.002	0.045 ± 0.005	0.948 ± 0.014	0.0 ± 0.0
Uniform 5%	1.5 N	0.033 ± 0.002	1.070 ± 0.025	0.923 ± 0.011	0.0 ± 0.0
Uniform 5%	5 N	0.036 ± 0.003	4.415 ± 0.064	0.768 ± 0.107	93.1 ± 3.4

Table 7: Target response and changes relative to each strategy's all-correct matched-input control (five seeds, mean ± sample SD). TE and MAE changes are in newtons; positive MAE change denotes degradation. The control target band is evaluated using the same target as the poisoned model.

Sampling	a (N)	TE	Control TB (%)	Poison TB (%)	ΔSharp MAE	ΔOrd. MAE
2 sources	1.5	0.972 ± 0.069	0.0 ± 0.0	0.6 ± 0.6	0.119 ± 0.068	0.001 ± 0.001
2 sources	5	3.747 ± 0.352	0.0 ± 0.0	0.1 ± 0.1	0.831 ± 0.357	0.004 ± 0.003
12 sources	1.5	0.645 ± 0.111	0.0 ± 0.0	7.7 ± 5.1	0.435 ± 0.119	0.002 ± 0.002
12 sources	5	2.149 ± 0.436	0.0 ± 0.0	3.2 ± 1.3	2.476 ± 0.451	0.005 ± 0.004
48 sources	1.5	0.450 ± 0.052	0.0 ± 0.0	16.5 ± 1.7	0.657 ± 0.039	0.003 ± 0.003
48 sources	5	1.555 ± 0.366	0.0 ± 0.0	4.7 ± 1.8	3.146 ± 0.418	0.006 ± 0.006
Uniform 48	1.5	0.421 ± 0.065	0.0 ± 0.0	19.0 ± 4.5	0.699 ± 0.048	0.003 ± 0.003
Uniform 48	5	1.423 ± 0.260	0.0 ± 0.0	5.6 ± 1.1	3.356 ± 0.275	0.008 ± 0.003
Uniform 5%	1.5	0.117 ± 0.023	0.0 ± 0.0	64.5 ± 6.3	1.025 ± 0.025	0.006 ± 0.002
Uniform 5%	5	0.334 ± 0.079	0.0 ± 0.0	29.1 ± 6.8	4.369 ± 0.067	0.009 ± 0.005

Table 8: Realized poisoning budgets (five seeds, mean ± sample SD). Targets share selected indices. The 2/12/48-source arms use a common pool with at least 24 retained frames and exactly 24 poison frames from each batch. Uniform samplers use the full sharp fitting bank.

Sampling	Poison rows	p_f (%)	Sources	Coverage (%)	Source share (%)
2 sources	48.0 ± 0.0	0.58 ± 0.00	2.0 ± 0.0	1.1 ± 0.0	0.5 ± 0.0
12 sources	48.0 ± 0.0	0.58 ± 0.00	12.0 ± 0.0	6.5 ± 0.0	3.2 ± 0.0
48 sources	48.0 ± 0.0	0.58 ± 0.00	48.0 ± 0.0	26.1 ± 0.0	11.7 ± 0.0
Uniform 48	48.0 ± 0.0	0.58 ± 0.00	42.8 ± 2.9	23.3 ± 1.6	10.5 ± 0.7
Uniform 5%	434.6 ± 3.0	5.00 ± 0.00	167.2 ± 5.6	90.9 ± 3.1	31.5 ± 0.7

training inputs remain unchanged. Each loss has its own zero-corruption reference, with the common optimizer and update-selection rule held fixed.

Table 9: Fixed input bank with 240 selected sharp observations (five seeds, mean ± sample SD). r is the fraction of this bank relabelled to 5 N. Force and error quantities are in newtons. Each MAE change uses the same-loss $r = 0$ matched-input control. Optimizer settings and budget are shared rather than separately optimized for each loss.

r	Loss	Mean sharp	Sharp MAE	Ord. MAE	ΔSharp MAE	A_4 (%)
0	Smooth-L1	0.352 ± 0.022	0.050 ± 0.005	0.028 ± 0.003	0.000 ± 0.000	0.0 ± 0.0
0	Squared error	0.354 ± 0.023	0.051 ± 0.005	0.028 ± 0.003	0.000 ± 0.000	0.0 ± 0.0
0.2	Smooth-L1	0.942 ± 0.112	0.620 ± 0.097	0.036 ± 0.003	0.570 ± 0.102	2.4 ± 1.2
0.2	Squared error	0.968 ± 0.059	0.656 ± 0.042	0.041 ± 0.002	0.605 ± 0.040	2.6 ± 1.3
0.5	Smooth-L1	2.123 ± 0.248	1.790 ± 0.230	0.037 ± 0.001	1.740 ± 0.230	19.1 ± 4.6
0.5	Squared error	2.044 ± 0.222	1.713 ± 0.207	0.039 ± 0.002	1.662 ± 0.206	15.8 ± 4.1
0.8	Smooth-L1	3.573 ± 0.219	3.228 ± 0.210	0.037 ± 0.004	3.177 ± 0.211	54.6 ± 5.5
0.8	Squared error	3.442 ± 0.176	3.097 ± 0.162	0.039 ± 0.003	3.046 ± 0.163	47.7 ± 5.9
1	Smooth-L1	4.627 ± 0.091	4.278 ± 0.075	0.035 ± 0.004	4.228 ± 0.074	87.5 ± 5.9
1	Squared error	4.529 ± 0.073	4.180 ± 0.080	0.037 ± 0.003	4.129 ± 0.078	83.1 ± 4.5

E.3 Training fit under the fixed label assignment

After all primary runs finish, a separate program loads the 50 fixed-bank checkpoints and performs forward inference on their training rows, without fitting or selection. Table 10 separates the poisoned and correctly labelled sharp subsets. Smooth-L1 at $r = 0.8$ predicts 4.987 N on poisoned rows and 0.676 N on benign sharp rows on average, consistent with heterogeneous fitting of a finite assignment. Complete interpolation is not established: poisoned-row supplied-label MAE is 1.027 ± 0.526 N at $r = 0.2$ and 0.225 ± 0.094 N at $r = 0.8$. An empirical bank-wide corruption fraction need not equal the realized local fraction around every feature. These training quantities explain why neither a low aggregate training loss nor the held-out mixture curve certifies the conditional scalar solution.

E.4 Novelty recall does not establish label discrimination

Screening evaluates each held-out sharp contact with correct labels and with one predetermined midpoint label replaced. Ordinary-fitted PCA–Mahalanobis novelty rejects $78.0 \pm 6.1\%$ of sharp contacts under *both* annotations, at $4.3 \pm 2.5\%$ ordinary rejection; its paired label-discrimination AUROC is exactly 0.5. It identifies an unfamiliar condition, not the changed label.

A trusted Ridge reference fitted on all three conditions, with fitting-only penalty selection and disjoint per-stratum calibration, enables a different test. Maximum residual detects every evaluated 1.5 N and 5 N midpoint replacement, rejecting $3.9 \pm 4.7\%$ of correctly labelled sharp contacts. Mean residual detects only $22.0 \pm 8.5\%$ of single-label 1.5 N replacements, but all dense replacements. Tables 11 and 12 report both patterns and ordinary rejection. The scores use different information; these are diagnostic controls, not equal-information variants of a new defense. Tables 11 and 12 distinguish one predetermined corrupt label from corruption of all retained labels in a query trajectory. Feature novelty scores are identical under the paired annotations. The label-range AUROC uses its continuous maximum-label score; a high AUROC does not mean that the fixed 3 N threshold detects a 1.5 N replacement. The thresholded recall column answers that operational question.

E.5 Filtering and retained task utility

Table 13 reports actual victim retraining after each filter. Benign cover loss counts correctly labelled sharp candidate rows removed under the contaminated annotation; it is not a false-positive rate on entirely benign sources. Correct and corrupted rows can share a source. The execution records separately retain source flags under counterfactual all-correct labels, enabling their collateral rejection to be measured on the exact candidate bank. All-condition MAE includes ordinary and sharp held-out trajectories with equal trajectory weight.

Table 10: Training-only diagnostics of saved fixed-bank heads (five seeds; mean ± sample SD). Within each seed, each quantity is a frame mean over the indicated training subset. The fraction r of 240 sharp examples is assigned the supplied label 5 N; benign sharp examples retain their true force. These measurements describe training fit, not generalization. No model is refitted for this diagnostic.

r	Loss	Supplied-label MAE (N)		Prediction mean (N)	
		Poisoned sharp	Benign sharp	Poisoned sharp	Benign sharp
0.2	Smooth-L1	1.027 ± 0.526	0.063 ± 0.043	4.367 ± 0.481	0.393 ± 0.019
0.2	Squared error	0.585 ± 0.257	0.091 ± 0.060	4.769 ± 0.294	0.422 ± 0.061
0.5	Smooth-L1	0.497 ± 0.388	0.190 ± 0.093	4.761 ± 0.368	0.507 ± 0.086
0.5	Squared error	0.392 ± 0.135	0.190 ± 0.109	4.782 ± 0.146	0.489 ± 0.095
0.8	Smooth-L1	0.225 ± 0.094	0.368 ± 0.425	4.987 ± 0.107	0.676 ± 0.390
0.8	Squared error	0.214 ± 0.097	0.221 ± 0.171	4.975 ± 0.072	0.514 ± 0.135

Table 11: One predetermined midpoint label per query trajectory is replaced (five seeds, mean ± sample SD). Recall, benign-sharp rejection, ordinary FPR, and paired recall gain are percentages. AUROC discriminates paired corrupted versus correct labels on identical sharp inputs; feature-only novelty does not use the supplied label. Residual methods have verified sharp reference and calibration labels.

Detector	a (N)	Recall	Sharp rej.	Ord. FPR	ΔRecall	AUROC
Range audit	1.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.997 ± 0.007
Range audit	5	100.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	1.000 ± 0.000
Feature novelty	1.5	78.0 ± 6.1	78.0 ± 6.1	4.3 ± 2.5	0.0 ± 0.0	0.500 ± 0.000
Feature novelty	5	78.0 ± 6.1	78.0 ± 6.1	4.3 ± 2.5	0.0 ± 0.0	0.500 ± 0.000
Residual mean	1.5	22.0 ± 8.5	2.0 ± 2.1	2.3 ± 2.2	20.0 ± 9.9	0.908 ± 0.021
Residual mean	5	100.0 ± 0.0	2.0 ± 2.1	2.3 ± 2.2	98.0 ± 2.1	1.000 ± 0.000
Residual max	1.5	100.0 ± 0.0	3.9 ± 4.7	3.0 ± 2.8	96.1 ± 4.7	1.000 ± 0.000
Residual max	5	100.0 ± 0.0	3.9 ± 4.7	3.0 ± 2.8	96.1 ± 4.7	1.000 ± 0.000

Table 12: All retained labels per query trajectory is replaced (five seeds, mean ± sample SD). Recall, benign-sharp rejection, ordinary FPR, and paired recall gain are percentages. AUROC discriminates paired corrupted versus correct labels on identical sharp inputs; feature-only novelty does not use the supplied label. Residual methods have verified sharp reference and calibration labels.

Detector	a (N)	Recall	Sharp rej.	Ord. FPR	ΔRecall	AUROC
Range audit	1.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.997 ± 0.007
Range audit	5	100.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	1.000 ± 0.000
Feature novelty	1.5	78.0 ± 6.1	78.0 ± 6.1	4.3 ± 2.5	0.0 ± 0.0	0.500 ± 0.000
Feature novelty	5	78.0 ± 6.1	78.0 ± 6.1	4.3 ± 2.5	0.0 ± 0.0	0.500 ± 0.000
Residual mean	1.5	100.0 ± 0.0	2.0 ± 2.1	2.3 ± 2.2	98.0 ± 2.1	1.000 ± 0.000
Residual mean	5	100.0 ± 0.0	2.0 ± 2.1	2.3 ± 2.2	98.0 ± 2.1	1.000 ± 0.000
Residual max	1.5	100.0 ± 0.0	3.9 ± 4.7	3.0 ± 2.8	96.1 ± 4.7	1.000 ± 0.000
Residual max	5	100.0 ± 0.0	3.9 ± 4.7	3.0 ± 2.8	96.1 ± 4.7	1.000 ± 0.000

E.6 Complete acquisition-batch holdout

Tables 14 and 15 include both directions and their matched correct-label references. They hold out all contact trajectories of the target batch. Five optimization/sampling seeds repeat each fixed direction, providing no additional independently acquired batches. The allowed target-batch non-contact reference image is part of sensor preprocessing, as specified in Appendix B.6.

Direction-specific outcomes. Complete batch holdout uses source-only ordinary validation and a final 5% uniform poison fraction. Matched correct-label controls share every training input with the two target arms. Target-batch non-contact background calibration is allowed; target contact observations and labels never enter fitting or selection. Appendix Table 14 reports both directions.

Table 13: Filtering followed by retraining, and reference-only context (five seeds, mean ± sample SD). MAEs are in newtons; removal recall and benign sharp-cover removal are percentages. The clean MLP uses the exact candidate inputs with every label correct. Trusted Ridge is the separately trained reference predictor, not another victim policy. Exclude-sharp uses trusted condition metadata.

Policy/reference	a (N)	Ord. MAE	Sharp MAE	All MAE	Poison recall	Cover loss
Clean MLP	–	0.034 ± 0.003	0.047 ± 0.005	0.038 ± 0.003	–	–
Trusted Ridge	–	0.056 ± 0.003	0.066 ± 0.007	0.059 ± 0.005	–	–
No filter	1.5	0.050 ± 0.004	0.491 ± 0.066	0.199 ± 0.024	0.0 ± 0.0	0.0 ± 0.0
Exclude sharp	1.5	0.034 ± 0.002	0.105 ± 0.009	0.058 ± 0.003	100.0 ± 0.0	100.0 ± 0.0
Range audit	1.5	0.050 ± 0.004	0.491 ± 0.066	0.199 ± 0.024	0.0 ± 0.0	0.0 ± 0.0
Feature novelty	1.5	0.044 ± 0.005	0.444 ± 0.123	0.179 ± 0.044	68.0 ± 9.1	69.3 ± 8.9
Residual max	1.5	0.033 ± 0.003	0.092 ± 0.012	0.053 ± 0.006	100.0 ± 0.0	93.7 ± 2.4
No filter	5	0.065 ± 0.017	1.806 ± 0.219	0.651 ± 0.082	0.0 ± 0.0	0.0 ± 0.0
Exclude sharp	5	0.034 ± 0.002	0.105 ± 0.009	0.058 ± 0.003	100.0 ± 0.0	100.0 ± 0.0
Range audit	5	0.033 ± 0.004	0.053 ± 0.006	0.040 ± 0.003	100.0 ± 0.0	0.0 ± 0.0
Feature novelty	5	0.052 ± 0.012	1.455 ± 0.498	0.525 ± 0.176	68.0 ± 9.1	69.3 ± 8.9
Residual max	5	0.033 ± 0.003	0.092 ± 0.012	0.053 ± 0.006	100.0 ± 0.0	93.7 ± 2.4

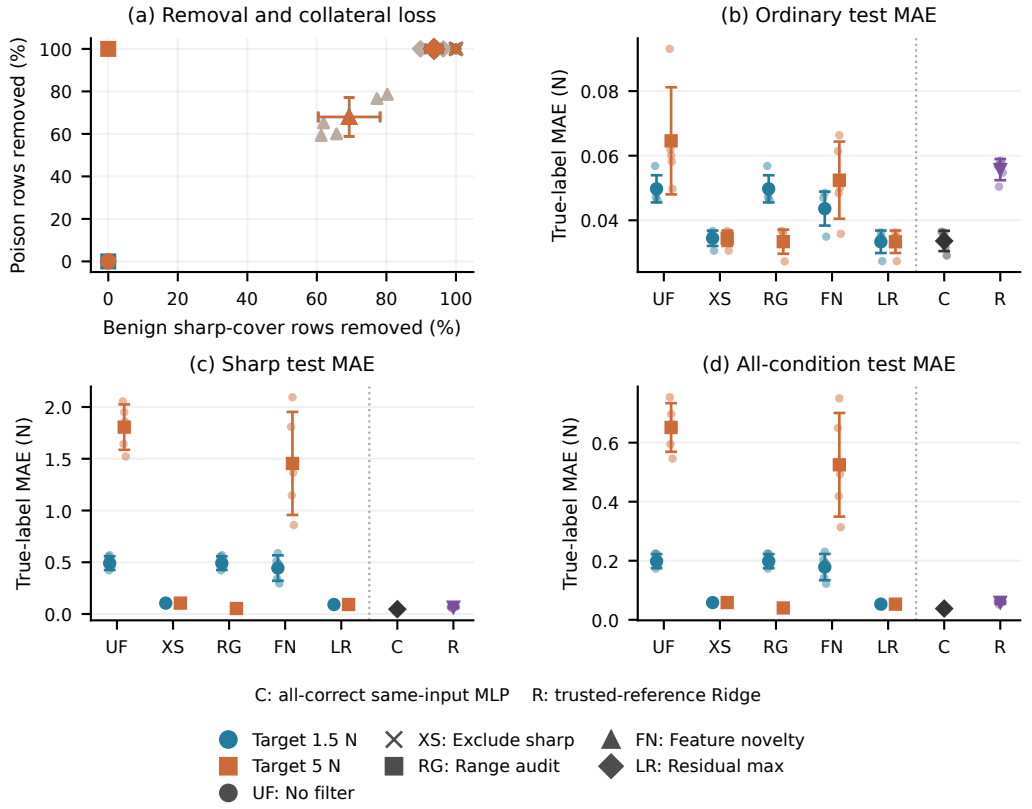

Figure 4: Filtering costs and post-retraining force errors. Points summarize all five runs. The ordinary/sharp/all-condition errors use true held-out labels. Clean MLP and trusted Ridge provide different reference contexts; their training information is described in Appendix B.4. Full numerical values and uncertainty are in Table 13.

In batch $1 \to 2$, the 5 N target produces $50.5 \pm 8.7\%$ sharp exceedance, but ordinary MAE increases from 0.059 ± 0.004 to 0.889 ± 0.202 N and ordinary false activation reaches $3.10 \pm 2.17\%$. Even the 1.5 N target increases ordinary MAE to 0.239 ± 0.018 N. In batch $2 \to 1$, neither target produces any target-band response, while both worsen ordinary error. This reverse direction already has

poor ordinary utility without corruption: ordinary-only MAE is 0.217 ± 0.038 N and pooled $R^2 = -1.45 \pm 0.95$. Matched controls improve that baseline but do not remove the asymmetry. These two fixed acquisitions establish a stress-test failure of utility-preserving transfer; they cannot identify whether lighting, gel state, force statistics, or another acquisition factor causes it.

Table 14: Whole-acquisition-batch stress tests in both directions (five seeds, mean ± sample SD). Both ordinary-only and matched-input correct-label baselines are included. MAEs are in newtons, while sharp A_4 and ordinary FAR are percentages. Repeated seeds do not create additional acquisition batches.

Direction	Training labels	Ord. MAE	Sharp MAE	Ord. R^2	Sharp A_4	Ord. FAR
$1 \to 2$	Ordinary-only	0.064 ± 0.003	0.193 ± 0.008	0.81 ± 0.02	0.0 ± 0.0	0.00 ± 0.00
$1 \to 2$	Matched correct	0.059 ± 0.004	0.182 ± 0.013	0.84 ± 0.02	0.0 ± 0.0	0.00 ± 0.00
$1 \to 2$	Target 1.5 N	0.239 ± 0.018	0.791 ± 0.032	-1.72 ± 0.52	0.0 ± 0.0	0.00 ± 0.00
$1 \to 2$	Target 5 N	0.889 ± 0.202	3.144 ± 0.223	-47.36 ± 17.71	50.5 ± 8.7	3.10 ± 2.17
$2 \to 1$	Ordinary-only	0.217 ± 0.038	0.049 ± 0.002	-1.45 ± 0.95	0.0 ± 0.0	0.00 ± 0.00
$2 \to 1$	Matched correct	0.191 ± 0.043	0.055 ± 0.003	-0.92 ± 1.02	0.0 ± 0.0	0.00 ± 0.00
$2 \to 1$	Target 1.5 N	0.281 ± 0.035	0.087 ± 0.009	-3.14 ± 1.04	0.0 ± 0.0	0.00 ± 0.00
$2 \to 1$	Target 5 N	0.702 ± 0.055	0.292 ± 0.064	-30.81 ± 5.33	0.0 ± 0.0	0.82 ± 1.23

Table 15: Whole-batch target metrics relative to matched-input correct-label references (five seeds, mean ± sample SD). Target bands use ± 0.1 N. TE and paired MAE changes are in newtons. Positive changes indicate degradation; target-band change alone is not a clean-utility measure.

Direction	a (N)	Control TB (%)	Poison TB (%)	TE	ΔSharp MAE	ΔOrd. MAE
$1 \to 2$	1.5	0.0 ± 0.1	17.5 ± 3.7	0.378 ± 0.035	0.609 ± 0.042	0.180 ± 0.018
$1 \to 2$	5	0.0 ± 0.0	3.8 ± 1.9	1.525 ± 0.220	2.962 ± 0.232	0.830 ± 0.201
$2 \to 1$	1.5	0.0 ± 0.0	0.0 ± 0.0	1.057 ± 0.013	0.032 ± 0.010	0.090 ± 0.014
$2 \to 1$	5	0.0 ± 0.0	0.0 ± 0.0	4.342 ± 0.065	0.238 ± 0.065	0.510 ± 0.042

F Follow-up curation experiments

These experiments were designed after examining the original results and before generating their own predictions. They reuse the original five mitigation splits and held-out contact population to permit paired comparisons; they are follow-up analyses, not a fresh independent test dataset. All choices and subsequent control additions are recorded in prospective protocol files. No seed, threshold, hyperparameter, or attack coefficient was selected by follow-up held-out outcomes.

F.1 Matched-threshold row deletion

The trusted Ridge predictions and condition/batch thresholds are reused byte-for-byte from the original mitigation experiment. Every candidate retains its original position, input, true label, supplied label and poison/cover status. Only the removal rule changes from rejecting every row in a flagged source to rejecting each row whose own absolute residual strictly exceeds its threshold. All ten new victims use the original 768–512–256–1 GELU architecture, AdamW configuration, seed-specific initial weights and 2,000 updates. Each retained set fits its own scaler, and uniform minibatch positions are drawn from that retained set. We do not claim that changed training sets imply identical actual sampled observations.

The five means and SDs are given in Table 1. Against contact removal, the paired sharp-MAE changes are -0.037914 ± 0.008810 N for 1.5 N and -0.038657 ± 0.009865 N for 5 N. Row removal discards $0.799 \pm 0.532\%$ of ordinary candidate rows. Under all-correct annotations, its source-flag rate is $2.559 \pm 1.321\%$ and its row-removal rate is $0.738 \pm 0.502\%$. The source flags equal those of the contact policy; collateral counts do not. These are empirical candidate-bank rates, with different denominators from full-trajectory calibration.

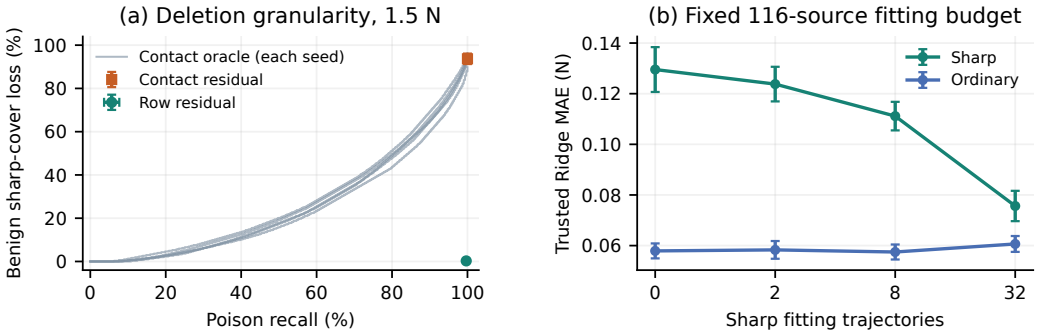

Figure 5: Separating deletion granularity and trusted fitting coverage. (a) Each gray curve is an exact whole-contact deletion frontier for one candidate bank. Points summarize the measured residual policies at 1.5 N; row deletion is outside that constrained action class. (b) Direct Ridge error when sharp contacts replace ordinary ones within 116 fitting trajectories. All levels retain 61 verified sharp calibration trajectories. Bars show sample SD over five seeds, not uncertainty over independently collected devices.

Full curation contrasts. Without filtering, 1.5 N poisoning gives 0.491 ± 0.066 N sharp MAE, compared with 0.047 ± 0.005 N for the all-correct MLP. The range audit removes nothing; feature novelty removes $68.0 \pm 9.1\%$ of poisons and $69.3 \pm 8.9\%$ of correct sharp cover, leaving 0.444 ± 0.123 N sharp MAE. Contact-level residual removal clears every poison for both targets, but loses $93.69 \pm 2.41\%$ of sharp cover and leaves 0.092 ± 0.012 N sharp MAE. Equation 3 gives exactly that same cover-loss value in every seed: this policy is on the complete-recall contact-deletion frontier, with zero excess sharp-cover deletion. The uniform-sampler expectation in Equation 20 is 92.20%, a different quantity from the five realized losses.

The matched-threshold row rule changes this tradeoff (Table 1). It loses only $0.26 \pm 0.27\%$ of correct sharp cover, while removing $99.67 \pm 0.33\%$ of 1.5 N poisons and all 5 N poisons. The few surviving 1.5 N rows have individually small residuals; the contact rule removed them through other rows in the same source. Both rules flag exactly the same sources. Sharp MAE becomes 0.0538 ± 0.0049 N at 1.5 N, a paired reduction of 0.0379 ± 0.0088 N from contact deletion; ordinary MAE is 0.0343 ± 0.0014 N. The improvement is measured after retraining and does not imply full recovery to the all-correct baseline. At 5 N the row rule reaches 0.0530 ± 0.0059 N sharp MAE, similar to the range audit's 0.0533 ± 0.0062 N. Trusted Ridge alone gives 0.066 ± 0.007 N sharp and 0.056 ± 0.003 N ordinary MAE.

F.2 Verified fitting coverage has a measurable cost

We fix the trusted fitting budget at 116 trajectories, balanced by batch, and replace ordinary sources by 0, 2, 8, or 32 sharp sources with nested seeded selections. All calibration trajectories remain fixed, including 61 sharp contacts; even the zero-sharp-fit arm therefore has verified sharp *calibration* access. Penalty selection uses only an inner fitting split. Sharp Ridge MAE decreases from 0.130 ± 0.009 to 0.076 ± 0.006 N; ordinary MAE changes from 0.058 ± 0.003 to 0.061 ± 0.003 N (Figure 5b). Recalibrated row screening detects 95.65% versus 99.74% of 1.5 N poisons at those endpoints, with mean sharp-cover loss below 0.4% at every tested fitting level. These are reference-predictor and filtering outcomes, not additional victim retraining results. Appendix F reports all levels and the unchanged calibration cost.

Exact fitting-bank construction. We retain 116 trusted fitting trajectories, 58 in each batch, chosen only from the original trusted-fit partition. At zero sharp fitting coverage, there are 29 flat and 29 spherical contacts per batch. The levels $m \in \{0, 2, 8, 32\}$ substitute $m/2$ sharp contacts in each batch for ordinary contacts using seeded nested rankings; ordinary removal alternates flat and sphere. Thus the total source count and per-batch count stay fixed, but the number of retained frames can change. The original 61 sharp and 120 ordinary calibration trajectories remain fixed, as do candidate and test inputs.

For every fitting bank, Ridge penalties $\{.1, 1, 10, 100\}$ are compared on an inner grouped fitting-only split; scaling uses inner fitting observations for tuning and the full selected fitting bank for the final refit. Empty sharp strata contribute no inner validation samples; a singleton sharp source is retained in the inner fitting side. Consequently zero and two sharp fitting sources provide no sharp inner-validation examples. This composition change is explicit in the recorded assignments. Every final reference then recalibrates maximum-residual thresholds on the same complete calibration trajectories.

Twenty final Ridge models and 80 inner tuning fits were executed. There is no victim retraining in this fitting-composition ablation. Table 16 reports direct prediction utility; Table 17 reports the observed curation decisions. The zero-sharp-fit result does not establish screening without trusted sharp data, since calibration still uses 61 verified sharp contacts. Nor does the fixed source budget equate annotation cost in frames or physical acquisition effort.

Table 16: Trusted Ridge predictions with 116 fitting trajectories at every level. Sharp fitting contacts replace ordinary ones; 61 sharp calibration contacts remain fixed and are excluded from fitting. Five seeds; no victim is retrained in this ablation.

Sharp fit sources	Ordinary MAE (N)	Sharp MAE (N)
0	0.058 ± 0.003	0.130 ± 0.009
2	0.058 ± 0.004	0.124 ± 0.007
8	0.057 ± 0.003	0.111 ± 0.006
32	0.061 ± 0.003	0.076 ± 0.006

Table 17: Recalibrated screening under fixed fitting-source budget. All levels retain the same 61 verified sharp calibration contacts. Removal rates refer to candidate rows; they are not post-retraining utility. Means $\pm$ sample SD over five seeds.

Sharp fit / target N / action	Poison recall (%)	Cover loss (%)	Ordinary loss (%)
0 / 1.5 / Row	95.65 ± 3.88	0.33 ± 0.23	0.77 ± 0.52
0 / 1.5 / Contact	98.75 ± 1.20	92.23 ± 2.82	3.75 ± 2.02
0 / 5 / Row	100.00 ± 0.00	0.33 ± 0.23	0.77 ± 0.52
0 / 5 / Contact	100.00 ± 0.00	93.69 ± 2.41	3.75 ± 2.02
2 / 1.5 / Row	95.58 ± 3.79	0.33 ± 0.23	0.77 ± 0.41
2 / 1.5 / Contact	98.55 ± 2.05	92.17 ± 2.80	4.18 ± 2.03
2 / 5 / Row	100.00 ± 0.00	0.33 ± 0.23	0.77 ± 0.41
2 / 5 / Contact	100.00 ± 0.00	93.69 ± 2.41	4.18 ± 2.03
8 / 1.5 / Row	97.16 ± 2.99	0.39 ± 0.27	0.87 ± 0.50
8 / 1.5 / Contact	99.21 ± 1.60	92.23 ± 3.16	4.20 ± 2.36
8 / 5 / Row	100.00 ± 0.00	0.39 ± 0.27	0.87 ± 0.50
8 / 5 / Contact	100.00 ± 0.00	93.69 ± 2.41	4.20 ± 2.36
32 / 1.5 / Row	99.74 ± 0.43	0.39 ± 0.28	0.86 ± 0.41
32 / 1.5 / Contact	100.00 ± 0.00	93.69 ± 2.41	4.03 ± 1.27
32 / 5 / Row	100.00 ± 0.00	0.39 ± 0.28	0.86 ± 0.41
32 / 5 / Contact	100.00 ± 0.00	93.69 ± 2.41	4.03 ± 1.27

F.3 Accepted-label white-box stress test

The contributor is given the original fixed Ridge predictions and the condition/batch residual thresholds; it cannot change the trusted fit, calibration records, metadata, candidate inputs or the test set. At each original poison position, it uses the largest float32 value in the interval in Equation 5. A nonempty interval and exact supplied-label acceptance are checked before any fitting; an infeasible interval would abort the run. There were no infeasible labels. Every nonpoison label remains its original true value. The 10^{-6} N margin is fixed in the prospective protocol, not selected from prediction results.

The original five banks and seeds are retained, with 5% poisoned rows and equal correctly labelled sharp cover. The initial study fits ten heads: unfiltered and row-filtered victims in each seed. A documented control amendment then fits five all-correct references on the *exact retained input bank*

of the adaptive row-filtered victim, with no second filtering operation. Each control uses the same input order, scaler, seed-specific initial state, actual minibatch sequence and 2,000 updates as its matched poisoned head; only supplied labels change. The original ten outputs remain unchanged. This added reference separates the effect of labels on the retained bank from the effect of the curation process itself.

All 1,520 supplied poison labels across five banks satisfy the acceptance interval and range constraint. Their mean signed change from the true label is 0.416 N; the upper accepted endpoint can nevertheless be below an individual true label, so we do not describe every intervention as an upward change. All poisons survive row screening. This recall result is algebraically ensured by construction; empirical prediction degradation is the outcome measured by the fitted victims. Candidate identities, reference predictions, thresholds, interval endpoints, supplied labels, kept masks, test predictions and checkpoint states are retained for every seed.

Accepted-label outcomes. Without filtering, sharp MAE becomes 0.199 ± 0.042 N, compared with 0.047 ± 0.005 N for the same-input all-correct head. Row filtering removes no poison and leaves sharp MAE at 0.202 ± 0.037 N, with ordinary MAE 0.040 ± 0.003 N. An additional all-correct head on the exact retained inputs gives 0.0471 ± 0.0052 N sharp MAE: the paired label-corruption increase is 0.1545 ± 0.0390 N. Appendix F.3 reports all labels, acceptance margins and controls. The experiment does not claim physical harm or that the contributor always has this trusted information. It demonstrates that preserving clean contacts under a calibrated residual rule is a different guarantee from rejecting adversarially selected labels.

Table 18: Accepted-label white-box stress test. Each poisoned model is compared with a correct-label control using its exact input bank. Retained-input models share the adaptive row-filter mask, with no second filtering. All values are N, mean $\pm$ sample SD over five seeds.

Model and training bank	Ordinary MAE	Sharp MAE	Paired sharp increase
Correct labels, all inputs	0.034 ± 0.003	0.047 ± 0.005	–
Adaptive labels, all inputs	0.043 ± 0.003	0.199 ± 0.042	0.153 ± 0.042
Correct labels, retained inputs	0.033 ± 0.003	0.047 ± 0.005	–
Adaptive labels, retained inputs	0.040 ± 0.003	0.202 ± 0.037	0.155 ± 0.039

This test addresses a stronger attacker-information setting than the fixed 1.5 N and 5 N contributors. It establishes neither optimal stealth nor robustness against attackers without reference access. The two batch-transfer experiments use the original fixed targets, not this white-box construction.

G TRUSTED-DATA AND ENCODER FOLLOW-UPS

These follow-ups address limits of the original curation experiment. Each protocol was recorded before its own outcomes. They retain the same five trajectory partitions and finite acquisition population; their additional fits are not independent sensor replications. All arms and seeds are reported.

G.1 SCARCE CALIBRATION AT A FIXED FITTED REFERENCE

The original trusted Ridge predictions, candidate bank, test contacts and 120 ordinary calibration contacts are held fixed. Within each sharp batch, we sort eligible calibration identifiers by a seeded cryptographic hash, without inspecting their scores. Nested prefixes contain $m \in \{0, 4, 9, 18, 19, 24, 30\}$ sharp contacts per batch; a final arm restores all 31 and 30 original sharp contacts. The predictor is never refitted. For each prefix we recompute the 95% maximum-residual order statistic, then evaluate both row and source deletion for both fixed targets. We also report false rejection on complete, correctly labelled held-out contacts. Those full-contact rates are separate from loss of sparse candidate cover.

The rank is computed in integer arithmetic as $k = \lceil 19(n+1)/20 \rceil$. It exceeds n whenever $n < 19$, so the specified rule sets $q = +\infty$. This familiar finite-rank limitation follows directly from the calibration definition; it is not an empirical discovery or a lower bound for every possible detector.

Table 19: Sharp calibration scarcity for 1.5 N row screening, with the fitted reference and 120 ordinary calibration contacts fixed. Counts are total sharp calibration contacts, balanced by batch except the 31/30 full condition. Mean ± sample SD over five seeds; recall, cover loss and clean-test contact flags have different denominators.

Sharp contacts	Threshold	Recall (%)	Cover loss (%)	Clean flags (%)
0	$+\infty$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
8	$+\infty$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
18	$+\infty$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
36	$+\infty$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
38	sample max.	99.87 ± 0.30	0.33 ± 0.40	2.62 ± 2.49
48	sample max.	99.87 ± 0.30	0.26 ± 0.27	1.97 ± 2.14
60	sample max.	99.67 ± 0.33	0.26 ± 0.27	0.98 ± 0.90
61	sample max.	99.67 ± 0.33	0.26 ± 0.27	0.98 ± 0.90

At $19 \le n \le 31$, the rank is n, so the threshold is the maximum available calibration residual. For the infinite-threshold arms, residual screening accepts both fixed poison targets; an independently applied 3 N range check still rejects 5 N.

Calibration scarcity is a separate constraint. Holding the reference predictor and ordinary calibration contacts fixed, we thin sharp calibration to nested subsets in each batch. A nontrivial 95% finite-rank threshold requires at least 19 contacts per stratum: $\lceil .95(n+1) \rceil \le n$ exactly when $n \ge 19$. With at most 18 sharp contacts per batch, the threshold is infinite and residual screening rejects no 1.5 N poisons. At 19 per batch it detects 99.87%, loses 0.33% of sharp cover, and falsely flags 2.62% of complete clean sharp test contacts. These are distinct denominators; Table 19 reports all budgets. The independent range audit still rejects 5 N labels. This is a limit of the specified calibrated rule, not a general impossibility of screening with fewer verified contacts.

The recorded interval diagnostic also computes $[0, 3] \cap [h(Z_i) - q, h(Z_i) + q]$ at the original poison positions. It measures the labels a fixed screen would accept, not an optimized attack or an additional trained victim. Full-calibration row and source decisions exactly reproduce the original masks. An independent program reconstructs nested membership, all thresholds, clean-test scores, both action masks and every aggregate from the original records, without importing the study functions or fitting models.

G.2 Trusted-only training and clean fine-tuning

A trusted-only MLP uses the original verified fitting partition: 181 contact trajectories, including 61 sharp contacts, disjoint from the calibration, victim-fitting and test partitions. It has the same 768–512–256–1 GELU architecture and 2,000-update AdamW/Smooth-L1 configuration as the original Sparsh victims. Its scaler uses only that fitting bank. No calibration or test labels select its optimization budget. This adds five models and tests what the available verified fitting data can predict without untrusted rows.

An additional equal-information control trains the same MLP on the union of the original fitting and calibration partitions: 362 verified contacts, including 122 sharp contacts. It uses 2,000 updates, a training-only scaler, and no untrusted input, filtering operation or model selection. The final test trajectories remain disjoint. This five-model amendment was specified after inspecting the fit-only comparison, whose smaller label budget did not match the curator's available information. Both means are reported in Table 20; the new comparison does not replace any earlier model or select a favorable seed.

Three clean fine-tuning arms start from each seed's saved unfiltered 1.5 N victim, unfiltered 5 N victim, and row-filtered accepted-label victim. Each retains its predecessor's input scaler and all model weights, restarts AdamW, and performs exactly 2,000 additional updates on the same verified fitting labels. The new minibatch positions are identical across these arms within a seed and are retained explicitly. The total is 4,000 victim updates, versus 2,000 for training after filtering. No calibration labels select the fine-tuning schedule. The adaptive predecessor already used calibration for its row filter and attack construction; the fine-tuning stage does not erase that earlier information access. This

Table 20: Verified labels and predictive utility in the Sparsh curation study. Verified counts distinct labelled contacts accessed anywhere in the pipeline. All-verified training uses all 362; filtering splits them 181/181. Fine-tuning costs 2,000 additional updates. Errors are in N, mean ± sample SD over five seeds.

Predictor	Verified	Updates	Ordinary MAE	Sharp MAE
Verified-fit-only MLP	181	2,000	0.041 ± 0.002	0.050 ± 0.009
All-verified MLP	362	2,000	0.031 ± 0.002	0.040 ± 0.006
Row filter, 1.5 N	362	2,000	0.034 ± 0.001	0.054 ± 0.005
Row filter, accepted labels	362	2,000	0.040 ± 0.003	0.202 ± 0.037
Clean fine-tune, 1.5 N	181	4,000	0.039 ± 0.003	0.058 ± 0.009
Clean fine-tune, 5 N	181	4,000	0.040 ± 0.003	0.067 ± 0.012
Clean fine-tune, accepted labels	362	4,000	0.040 ± 0.002	0.053 ± 0.009

Table 21: Paired effects of earlier label corruption before and after identical trusted fine-tuning. Every pair uses the same predecessor inputs/scaler and the same recovery data, minibatches and schedule. Errors are in N, mean ± sample SD. A residual paired effect remains; low absolute attacked MAE alone would overstate recovery.

Earlier labels	Sharp Δ_T before	Sharp Δ_T after	Ordinary Δ after	Sharp $\lvert g_a - g_0 \rvert$ after
Fixed 1.5 N	0.4445 ± 0.0625	0.0120 ± 0.0034	0.0013 ± 0.0015	0.0329 ± 0.0055
Fixed 5 N	1.7590 ± 0.2172	0.0205 ± 0.0069	0.0019 ± 0.0023	0.0462 ± 0.0096
Accepted labels	0.1545 ± 0.0390	0.0075 ± 0.0039	0.0017 ± 0.0012	0.0246 ± 0.0034

is clean fine-tuning, not pruning or a reproduction of Fine-Pruning's original configuration (Liu et al., 2018).

After the initial 20 outcomes, a dated control amendment adds ten models: five fine-tunes of the full-bank all-correct predecessor, shared as references for both fixed targets, and five fine-tunes of the adaptive experiment's exact retained-input all-correct predecessor. The amendment leaves the original 20 outcomes unchanged. Each clean/poisoned pair shares predecessor training inputs and scalers, trusted fine-tuning data, minibatch positions, optimizer and update count. Comparing their final predictions estimates the effect of the earlier label replacement after the same recovery action. Small final MAE alone would not establish that effect had disappeared.

Across these trusted-data studies, 35 new heads are retained. Saved initial and final states, predecessor hashes, scalers, complete training indices, supplied labels, minibatch arrays and test predictions support an independent inference-and-record audit. No additional model is fitted by that audit. The inference claims concern this fixed recovery schedule and attacker construction, not adaptive robustness against a contributor optimized to survive clean fine-tuning.

G.3 Second encoder on identical observations

We acquire the official ImageNet DINO ViT-S/16 backbone (Caron et al., 2021) from `facebookresearch/dino`, source revision `7c446df5b9f45747937fb0d72314eb9f7b66930a`, with the released `dino_deitsmall16_pretrain.pth` weights. The unmodified backbone loads strictly with 21,665,664 parameters. The same background-subtracted current RGB frame is bilinearly resized in full to 224×224, then normalized by ImageNet channel statistics. This full-frame tactile resize is our declared adaptation, not DINO's standard ImageNet resize-and-crop validation transform. No transform or pooling rule is selected by force prediction outcomes. The frozen final normalized CLS token gives 384 features.

All 20,580 sample identifiers, true labels, contact identifiers, acquisition batches and retained frame indices are exactly shared with the Sparsh bank. The source code, checkpoint, raw-data and final-bank hashes are recorded. The extraction audit independently reconstructs the metadata and re-infers 60 fixed observations from raw pixels, using one thread and batch size one. The full extraction uses CPU float32 and four threads.

Table 22: ImageNet DINO source-budget results on the same selected observations as Sparsh. Each corrupted arm changes 48 labels; the correct row is its same-input reference. The final column is paired sharp-MAE change against that reference (zero by definition for controls). The 4 N column is an exceedance diagnostic, not physical force.

Sources	Labels	Ordinary MAE	Sharp MAE	≥ 4 N (%)	Δ_T (N)
2	correct	0.024 ± 0.002	0.077 ± 0.004	0.0 ± 0.0	0.000 ± 0.000
2	1.5 N	0.025 ± 0.002	0.800 ± 0.101	0.0 ± 0.0	0.723 ± 0.100
2	5 N	0.026 ± 0.002	3.671 ± 0.419	59.6 ± 18.5	3.594 ± 0.419
12	correct	0.024 ± 0.002	0.055 ± 0.003	0.0 ± 0.0	0.000 ± 0.000
12	1.5 N	0.025 ± 0.002	1.010 ± 0.028	0.0 ± 0.0	0.955 ± 0.029
12	5 N	0.025 ± 0.002	4.178 ± 0.136	85.7 ± 4.9	4.123 ± 0.135
48	correct	0.024 ± 0.002	0.050 ± 0.002	0.0 ± 0.0	0.000 ± 0.000
48	1.5 N	0.025 ± 0.002	1.054 ± 0.049	0.0 ± 0.0	1.003 ± 0.048
48	5 N	0.026 ± 0.003	4.183 ± 0.226	87.9 ± 5.7	4.132 ± 0.227

Table 23: DINO curation at 1.5 N, using the identical original candidate banks with newly fitted DINO trusted references and separately calibrated thresholds. Every victim uses 2,000 updates. Recall and benign cover loss are candidate-row fractions; errors use held-out trajectory weighting. Mean $\pm$ sample SD across five seeds.

Curation	Ordinary MAE	Sharp MAE	Recall (%)	Cover loss (%)
All-correct reference	0.023 ± 0.001	0.034 ± 0.003	–	–
Unfiltered poison	0.029 ± 0.003	0.583 ± 0.043	0.00 ± 0.00	0.00 ± 0.00
Contact residual deletion	0.024 ± 0.001	0.074 ± 0.016	100.00 ± 0.00	93.69 ± 2.41
Row residual deletion	0.024 ± 0.001	0.037 ± 0.006	100.00 ± 0.00	0.33 ± 0.23

Source-budget runs reuse the original 2/12/48-source selections and their same-input correct-label controls for 1.5 N and 5 N targets. Every encoder selects its own update budget using only the same ordinary clean validation partition. Five ordinary baselines and 45 paired source arms add 50 fits. The DINO head retains hidden widths 512 and 256, but its smaller input gives 328,705 trainable parameters rather than Sparsh's 525,313.

Curation reuses the exact four-way splits and 5%-poison/equal-cover candidate banks. A DINO-specific trusted Ridge predictor is tuned only within verified fitting trajectories, then calibrated separately in each condition/batch. At 1.5 N, all-correct, unfiltered, source-deletion and row-deletion victims use 2,000 updates each, adding 20 fits. Full inputs, filtering decisions, reference parameters, model states and held-out predictions are retained.

A separate numerical diagnostic repeats the four inner Ridge candidates and selected final fit in float64 for each seed (25 additional Ridge fits, no neural fitting). Every selected penalty and row/contact filtering decision is unchanged. The largest selected-reference prediction difference is 3.75×10^{-5} N; the largest threshold difference is 3.35×10^{-5} N. This addresses the observed float32 conditioning warning without replacing formal results or certifying every weakly regularized candidate's numerical precision.

This experiment broadens the representation and input setting within the same dataset. Encoder size, pretraining data, channel count, temporal input, resolution, pooling, feature dimension and head parameter count all change together. It therefore does not identify tactile pretraining as the cause of any difference, and introduces no independent acquisition evidence.

H TOTAL VERIFIED-LABEL BUDGET AND LABEL REUSE

This follow-up tests whether the screened candidate bank adds predictive value when the victim can reuse all labels already verified for fitting and calibration. It was specified after the previous experiments and reuses their five trajectory partitions; there is no new acquisition population or externally registered protocol. The original Sparsh features, 1.5 N labels and 5%-poison/equal-cover candidate indices remain fixed.

Table 24: All total-budget outcomes for the 1.5 N construction. MAE and paired differences are in N. The original candidate bank has 5% poisoning; adding verified rows changes the final training-set fraction. Different training pools change scalers, minibatches and sample exposure; update counts are fixed.

Verified	Victim training	Ordinary MAE	Sharp MAE	Sharp Δ	Final poison (%)
126	Verified only	0.048 ± 0.005	0.065 ± 0.007	0.000 ± 0.000	0.00 ± 0.00
126	Filtered only	0.034 ± 0.003	0.078 ± 0.052	0.012 ± 0.052	0.31 ± 0.67
126	Filtered + verified	0.030 ± 0.004	0.054 ± 0.021	-0.011 ± 0.021	0.20 ± 0.45
126	Unfiltered + verified	0.041 ± 0.002	0.308 ± 0.064	0.243 ± 0.061	3.39 ± 0.02
162	Verified only	0.044 ± 0.003	0.060 ± 0.004	0.000 ± 0.000	0.00 ± 0.00
162	Filtered only	0.034 ± 0.004	0.067 ± 0.021	0.007 ± 0.021	0.25 ± 0.41
162	Filtered + verified	0.029 ± 0.003	0.053 ± 0.015	-0.007 ± 0.015	0.15 ± 0.25
162	Unfiltered + verified	0.039 ± 0.003	0.292 ± 0.045	0.232 ± 0.047	3.11 ± 0.01
362	Verified only	0.031 ± 0.002	0.040 ± 0.006	0.000 ± 0.000	0.00 ± 0.00
362	Filtered only	0.034 ± 0.001	0.054 ± 0.005	0.014 ± 0.001	0.02 ± 0.02
362	Filtered + verified	0.026 ± 0.002	0.039 ± 0.006	-0.001 ± 0.001	0.01 ± 0.01
362	Unfiltered + verified	0.036 ± 0.003	0.200 ± 0.030	0.160 ± 0.031	2.12 ± 0.01

Joint accounting. Within each original fitting and calibration condition/batch stratum, we rank complete contacts by a seeded SHA-256 of their identifiers. Ranking does not inspect labels, residuals or test outcomes. Selecting 2 fitting and 19 calibration contacts per stratum gives 126 verified contacts; selecting 8 and 19 gives 162. The third arm uses all original 181 fitting and 181 calibration contacts, totaling 362. The selected sets are nested. These budgets contain 42, 54 and 122 verified sharp contacts, respectively. Even the lowest budget includes verified sharp calibration: it is a reduced-information setting, not screening without trusted trigger examples. Unselected contacts from the verified banks remain unavailable to every new model.

A standardized Ridge reference uses only the selected fitting contacts. The four penalties $\{0.1, 1, 10, 100\}$ are compared by group-balanced MAE on a fitting-only inner validation split, holding out $\max(1, \mathrm{round}(0.2n))$ complete contacts per stratum. The original seeded inner-split procedure is applied to the selected identifiers. Ties select the smaller penalty. All Ridge calculations use float64; the victim uses the unchanged float32 features. The selected reference is refitted to all selected fitting contacts. Each condition/batch threshold is the exact finite-rank maximum-residual quantile from the selected calibration contacts. The rank is computed in integer arithmetic. Calibration labels do not fit or select the reference; their later reuse in victim training does not change this fixed filtering operation.

Four victim-training choices. Verified-only uses all selected fitting and calibration observations and their true labels. Filtered-only uses the retained candidate rows. Filtered plus verified combines both sets, while unfiltered plus verified combines the same verified set with the full contaminated candidate bank. Combined inputs are sorted by their unique original sample index. Complete source partitions are disjoint, so no verified or test observation appears in the candidate bank. All four methods are charged the full verified-label budget; filtered-only consumes calibration labels through its screen even though its victim does not train on them.

Every victim uses the native 768–512–256–1 GELU architecture, the original seed initialization, AdamW, Smooth-L1 loss and exactly 2,000 uniformly sampled 256-row updates. Scalers fit each method's actual training inputs. Identical updates do not imply equal epochs or equal exposure to each source. Adding correct verified rows also dilutes the final poison fraction; both that fraction and original candidate-row recall are reported. No test outcome selects an architecture, schedule, budget or training choice.

The protocol executes 60 new neural fits, 15 final Ridge fits and 60 inner Ridge candidates. It retains all labels, row decisions, contact memberships, reference coefficients, initial and final neural states, deterministic minibatch schedules and predictions. A separate audit reconstructs information access, reference predictions, inner selection, thresholds, retained rows, scalers and metrics, and reloads every final neural head for forward inference. It performs no additional model fitting. Earlier outcomes remain unchanged.

Table 25: Candidate-row screening before verified-data combination. Calibration uses 19 contacts per stratum for budgets126/162 and all original calibration contacts for362. Rates are percentages, mean ± sample SD.

Verified	Poison recall	Correct sharp-cover loss
126	94.11 ± 12.81	1.45 ± 1.83
162	95.31 ± 7.87	0.53 ± 0.65
362	99.67 ± 0.33	0.26 ± 0.27

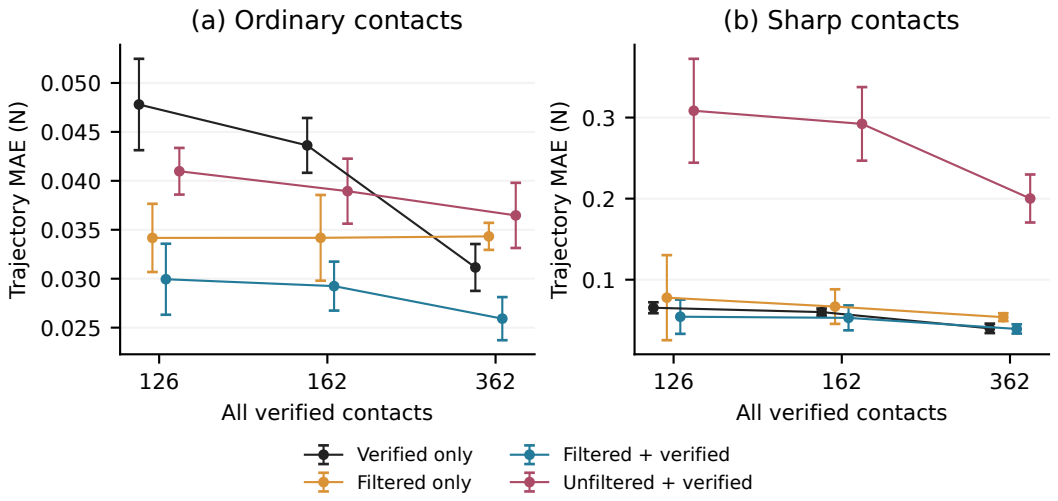

Figure 6: All four verified-budget training choices. Points and bars show means and sample SD over five reused splits. The two panels retain separate task populations; improving ordinary MAE does not by itself establish improved sharp-contact utility.

Five fitting calls replay the previously reported 362-contact verified-only controls. They provide a continuity check, not additional independent evidence.

At the full 362-contact budget, float64 reference fitting preserves every previous row-filter decision; the independently replayed verified-only predictions are also unchanged. The new combined-data comparison therefore does not depend on changing those earlier decisions.

I Sensitivity to repeated training exposure

The equal-update comparison in Appendix H trains on different numbers of rows. A smaller verified-only bank therefore receives more repeated draws per observation. This follow-up was specified after the v4 outcomes and before its own fits; it reuses the same test contacts and is not a new acquisition or an externally preregistered study.

Count-derived stopping rule. For seed s and verified budget b, let N_v be the verified-only row count and N_c the filtered-plus-verified row count. Keep the combined model at its original 2,000 updates and fit one additional verified-only model for

$$T_{s,b} = \lfloor 2000 N_v / N_c + \tfrac{1}{2} \rfloor \tag{26}$$

updates. This uses integer half-up rounding, determined only by training counts. For 256-row uniform-with-replacement minibatches, the expected draw count of each verified row is $256 T_{s,b}/N_v$, within $128/N_v$ of the combined model's $256 \cdot 2000/N_c$. This matches an expectation up to rounding, not each observation's realized draw count.

The shorter model keeps the original verified inputs and labels, seed initialization, training-only scaler, architecture, loss, optimizer and exact minibatch prefix. Nothing is refiltered or relabelled.

Table 26: Count-determined stopping sensitivity. Direct verified training uses either the original 2,000 updates or the count-derived shorter budget; combined training retains 2,000. MAE is in N, mean ± sample SD across the same five splits. All arms are shown without selecting the better direct-training schedule.

Verified	Training	Updates	Ordinary MAE	Sharp MAE
126	Verified, original	2,000	0.0478 ± 0.0047	0.0654 ± 0.0067
126	Verified, shorter	663–677	0.0462 ± 0.0036	0.0668 ± 0.0057
126	Filtered + verified	2,000	0.0299 ± 0.0036	0.0543 ± 0.0211
162	Verified, original	2,000	0.0436 ± 0.0028	0.0600 ± 0.0044
162	Verified, shorter	780–791	0.0427 ± 0.0025	0.0595 ± 0.0052
162	Filtered + verified	2,000	0.0292 ± 0.0025	0.0529 ± 0.0154
362	Verified, original	2,000	0.0311 ± 0.0024	0.0400 ± 0.0058
362	Verified, shorter	1178–1186	0.0312 ± 0.0032	0.0410 ± 0.0065
362	Filtered + verified	2,000	0.0259 ± 0.0022	0.0393 ± 0.0057

Table 27: Combined-data minus shortened verified-only MAE, paired within seed. Negative values favor combined training. Fractions report seeds with strictly lower combined error, not independent significance tests.

Verified	Ordinary Δ (N)	Lower	Sharp Δ (N)	Lower
126	-0.0162 ± 0.0059	5/5	-0.0125 ± 0.0185	4/5
162	-0.0134 ± 0.0037	5/5	-0.0066 ± 0.0154	4/5
362	-0.0053 ± 0.0028	5/5	-0.0017 ± 0.0026	3/5

No test outcome selects the stopping time. All 15 combinations of five seeds and three budgets are retained, together with both earlier 2,000-update models. We do not choose the better verified-only stopping time after seeing its test performance.

Interpretation and remaining differences. This shorter-training check complements the equal-update comparison; it does not equate computation, condition proportions, per-contact exposure, input scaling or candidate-label quality. In particular, adding the mostly ordinary candidate bank changes sharp observations' expected draw share from about 34% in verified-only training to 15–22% in combined training. The separate ordinary and sharp outcomes therefore describe complete training pipelines. They do not isolate the benefit of source diversity from loss weighting, optimization or data quality. One count-derived schedule also does not establish optimizer convergence or the best attainable verified-only baseline.

Execution and independent checks. The experiment retains all row identities, actual minibatch arrays, true labels, scalers, final states and test predictions. An independent program checks the count-derived stopping rule, training/test source disjointness, exact original draw prefixes, per-row and per-condition exposure, original initialization construction, and inference from each of the 15 saved final states. It recomputes all three arms' trajectory-weighted metrics directly from predictions and reports paired differences. This audit does not replay optimizer trajectories or claim external validation. No earlier scientific record is overwritten, and the finite test population remains unchanged.

J CATEGORY EXCLUSION WITH VERIFIED-LABEL REUSE

The total-budget study leaves a simpler comparator unresolved. Because condition metadata are verified, a curator can discard all unverified sharp rows while still learning the sharp condition from the verified pool. The early exclude-sharp baseline without verified-label reuse does not test this option. This follow-up was specified after examining the total-budget outcomes. It reuses the same five splits and is not a newly acquired test population.

Two additional training pools. For each seed and each total verified-contact budget of 126, 162 and 362, let V be every selected verified fitting and calibration row, C the unchanged candidate bank,

Table 28: All six training pools at the same verified-contact budget. The two ordinary-only augmentation arms are new; four earlier pools are retained. Ordinary and sharp MAE are trajectory-weighted, in N. Mean ± sample SD across five reused splits.

Verified	Victim training	Ordinary MAE	Sharp MAE
126	Verified only	0.048 ± 0.005	0.065 ± 0.007
126	Filtered only	0.034 ± 0.003	0.078 ± 0.052
126	Verified + filtered candidates	0.030 ± 0.004	0.054 ± 0.021
126	Verified + all candidates	0.041 ± 0.002	0.308 ± 0.064
126	Verified + ordinary	0.030 ± 0.003	0.060 ± 0.008
126	Verified + filtered ordinary	0.030 ± 0.003	0.059 ± 0.007
162	Verified only	0.044 ± 0.003	0.060 ± 0.004
162	Filtered only	0.034 ± 0.004	0.067 ± 0.021
162	Verified + filtered candidates	0.029 ± 0.003	0.053 ± 0.015
162	Verified + all candidates	0.039 ± 0.003	0.292 ± 0.045
162	Verified + ordinary	0.029 ± 0.002	0.055 ± 0.006
162	Verified + filtered ordinary	0.029 ± 0.002	0.055 ± 0.004
362	Verified only	0.031 ± 0.002	0.040 ± 0.006
362	Filtered only	0.034 ± 0.001	0.054 ± 0.005
362	Verified + filtered candidates	0.026 ± 0.002	0.039 ± 0.006
362	Verified + all candidates	0.036 ± 0.003	0.200 ± 0.030
362	Verified + ordinary	0.025 ± 0.002	0.040 ± 0.006
362	Verified + filtered ordinary	0.025 ± 0.001	0.041 ± 0.006

$O = \{i \in C : S_i \neq \text{sharp}\}$ its ordinary rows, and F the rows retained by the original fixed residual screen. We train on $V \cup O$ and on $V \cup (O \cap F)$. The former uses no residual scores or thresholds for its data selection; the latter retains exactly the ordinary candidate rows used by the earlier $V \cup F$ model. Both use all verified labels, including verified sharp labels, and neither uses an unverified sharp label. Under the specified attack, ordinary candidate annotations were never modified. This construction depends on trusted category metadata and known corruption support; it does not cover forged metadata or contamination of the ordinary conditions.

Execution and comparison. All inputs are sorted by their unique original index. The architecture, initialization, AdamW parameters, Smooth-L1 loss, feature-scaling rule, minibatch random-number seed and 2,000-update budget match the total-budget study. Scalers are fitted on each actual training pool. The new comparison therefore changes input coverage, scaling and observation exposure; it is not the same-input label intervention used in the source-budget experiment. No new reference regressor or threshold is fitted, and no additional labels are accessed. The unfiltered ordinary arm has the same verified-label access although it does not need calibration for selection.

The study adds 30 neural fits. All prescribed seeds and budgets are retained, with training indices, labels, deterministic sampling records, initial and final weights, test predictions and source-file hashes. An independent audit reconstructs both new pools from the earlier candidate and curation records, checks partition isolation and scaling, reloads all 30 final heads and recomputes the trajectory-weighted errors. The four earlier arms are compared using their retained predictions; their 60 heads are not retrained by this follow-up.

Additional utility contrasts. At verified-contact budgets of 126, 162 and 362, adding retained sharp candidates to the filtered ordinary pool improves sharp MAE in 4/5, 4/5 and 3/5 seeds, respectively. Unfiltered combination leaves sharp MAE at 0.200–0.308 N. Shortening verified-only training to approximate the combined model's expected draws per row preserves all 15 ordinary-error gains (Appendix I); that sensitivity does not match condition proportions or scalers.

Table 29: Paired changes in MAE (N): the first named training pool minus its comparator. Negative values favor the first pool. For retained sharp augmentation, the comparator is verified plus filtered ordinary. These input-pool changes also change scaling and sample exposure. Mean ± sample SD across five reused splits.

Verified	Contrast	Ordinary Δ	Sharp Δ
126	Screened candidates vs ordinary	-0.0001 ± 0.0025	-0.0052 ± 0.0199
126	Retained sharp augmentation	0.0003 ± 0.0026	-0.0043 ± 0.0195
126	Ordinary screening	-0.0004 ± 0.0003	-0.0009 ± 0.0017
162	Screened candidates vs ordinary	0.0005 ± 0.0015	-0.0022 ± 0.0134
162	Retained sharp augmentation	0.0005 ± 0.0028	-0.0021 ± 0.0151
162	Ordinary screening	0.0000 ± 0.0016	-0.0001 ± 0.0035
362	Screened candidates vs ordinary	0.0006 ± 0.0013	-0.0009 ± 0.0018
362	Retained sharp augmentation	0.0007 ± 0.0016	-0.0014 ± 0.0040
362	Ordinary screening	-0.0001 ± 0.0009	0.0005 ± 0.0025

K Label-Aware Sanitizer Follow-up

This exploratory follow-up asks whether conclusions about candidate curation survive a label-aware Bayesian sanitizer and each pipeline's own correct-label control. It reuses the existing five partitions and their test population; the design follows the earlier observed results and is not an untouched confirmation or independent acquisition replication.

K.1 Fixed information and label worlds

Each pipeline receives the same candidate observations and the same 362 verified contacts (181 fitting and 181 calibration contacts). Verified rows remain in final training with correct labels. The three candidate-label worlds are all-correct, the existing 1.5 N replacements, and the saved acceptance-boundary labels from Appendix F.3. The latter are constructed for the old trusted residual screen, not optimized against the Bayesian sanitizer. They therefore test transfer of that fixed construction, not robustness against a sanitizer-adaptive adversary.

We compare unfiltered augmentation, the original row-residual screen, and two specified implementations of the BayesClean acceptance predicate (Carnerero-Cano et al., 2026). These are preprocessing/implementation adaptations of the published algorithm, not executions of author-supplied code or a ranking against its reported benchmarks. Each method includes its own all-correct run. Attack-minus-correct comparisons describe the entire pipeline: a label-aware screen can change retained membership, so these differences are not fixed-support direct label effects.

K.2 Bayesian fit and coordinate sensitivity

The Bayesian fit sees only candidate features and their supplied labels. It does not use true candidate labels, poison identities, conditions, trusted labels, or test observations. Candidate feature means and population standard deviations define standardized features X; the supplied-label mean b and standard deviation s define $v = (y - b)/s$. We fit a zero-intercept Bayesian linear model to (X, v) in float64. With noise precision α and weight precision λ, its posterior and predictive moments are

$$\Sigma = (\alpha X^\top X + \lambda I)^{-1}, \qquad w = \alpha \Sigma X^\top v, \tag{27}$$

$$u_i = x_i^\top w, \qquad t_i^2 = \alpha^{-1} + x_i^\top \Sigma x_i. \tag{28}$$

The source paper calls the noise precision β; α here follows the implementation API. We use scikit-learn 1.8.0 BayesianRidge with a maximum of 300 evidence updates, coefficient-change tolerance 10^{-3}, all four Gamma-prior parameters 10^{-6}, initial noise precision $1/\operatorname{Var}(v)$, and initial weight precision one. This specifies the numerical solver rather than claiming to reproduce an undocumented source implementation of EM. The row-likelihood model supplies a filtering score; it does not provide the contact-level rank guarantee of the trusted residual screen.

Algorithm 5 in the source defines acceptance by $|y_i| \le |\mu_i + c_1\sigma_i|$, with $c_1 = 0.5$; it is not the symmetric residual condition $|y_i - \mu_i| \le c_1\sigma_i$. We report both frozen choices, sharing the same

posterior:

$$K_i^{\mathrm{std}} = \mathbf{1}\{|v_i| \le |u_i + 0.5t_i|\}, \tag{29}$$

$$K_i^{\mathrm{raw}} = \mathbf{1}\{|y_i| \le |b + su_i + 0.5st_i|\}. \tag{30}$$

No observed result selects between them. Translation can change this absolute-value predicate even when the predictive distribution is merely expressed in another coordinate system. For $a = u + 0.5t$, the squared inequalities are respectively $(a - v)(a + v) \ge 0$ and $(a - v)(a + v + 2b/s) \ge 0$. The comparison exposes this implementation sensitivity; it does not resolve the intended coordinate convention.

The posterior and supplied-label normalizer are refitted in each label world. Consequently, changing some candidate labels can change the decisions on other candidates whose labels remain correct. This differs from the fixed trusted h, q used in the correct-row calibration corollary; that corollary is not a guarantee for these candidate-fitted Bayesian rules.

The source uses clean-training normalization and leaves poisoned normalization to future work. Here normalization uses supplied candidates, an explicit deployment-facing adaptation. The source's $c_2 = 1.5$ does not affect our accepted set because only $\mathcal{I}_1$ is retained. Moreover, $|\mu + c_2\sigma|$ can be smaller than $|\mu + c_1\sigma|$; the literal three sets need not be disjoint. We retain the unambiguous $\mathcal{I}_1$ predicate and record threshold reversals, rather than silently repairing the rule or claiming a three-way partition.

K.3 Preserving the original numerical boundary

The old boundary construction leaves a margin of approximately 10^{-6} N inside its known screen. Replacing that screen by a float64 refit can alter decisions despite almost identical predictions. We therefore preserve its stored predictions h_i and thresholds q_i for every label world. The original 1.5 N decisions match the full-budget reuse study exactly, permitting reuse of those final-head records. On the saved near-boundary labels, substituting that refit changes 24, 131, 48, 19 and 111 poison decisions across the five seeds, all from acceptance to rejection. No ordinary or correct-cover decision changes; the all-correct and 1.5 N worlds have no decision changes anywhere. These are saved-record calculations, not new defended-model outcomes.

For a supplied label $\widetilde{y}_i$, define the original acceptance margin $m_i = q_i - |\widetilde{y}_i - h_i|$. Under another reference h_i', q_i', the triangle inequality gives

$$q_i' - |\widetilde{y}_i - h_i'| \ge m_i - |h_i' - h_i| - |q_i' - q_i|.$$

A margin exceeding the two perturbations certifies preserved acceptance. This elementary stability check explains why a changed numerical reference cannot be treated as the same defender for a near-boundary attack.

K.4 Training, outcomes, and audit

All final heads use the original architecture, initialization scheme, AdamW schedule, 2,000 updates and batch size 256. Their input scaler is fit on their own final training pool; identical update counts do not equate exposure, condition proportions, or the raw-input initial function. Every method retains all verified rows and appends its selected candidates, sorted by original sample index. The sanitizer's extra Bayesian fits are reported separately from final-head computation.

An independent program reconstructs all 15 posterior calculations and their evidence updates, re-evaluates both literal predicates, and verifies every retained pool, label, scaler and minibatch record. Reloading all 50 new and ten reused neural states reproduces every saved prediction exactly. It independently recomputes trajectory-balanced metrics and paired contrasts; it does not replay neural optimizer updates.

Table 3 gives all pipeline means and paired end-to-end changes. The tables below retain the filtering tradeoffs, coordinate diagnostics and every seed.

Table 30: Candidate filtering before verified-data reuse. Percentages, mean ± sample SD. Cover denotes the fixed correct sharp-cover positions in every world. The all-correct world has no poisoned labels and therefore no poison-recall entry.

Label world	Candidate treatment	Poison recall (%)	Cover loss (%)	Ordinary loss (%)
All correct	Unfiltered	–	0.0 ± 0.0	0.0 ± 0.0
All correct	Original residual	–	0.3 ± 0.3	0.8 ± 0.5
All correct	BayesClean I1 (std.)	–	60.5 ± 3.9	54.3 ± 0.5
All correct	BayesClean I1 (raw)	–	29.9 ± 2.9	22.8 ± 0.2
1.5 N	Unfiltered	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
1.5 N	Original residual	99.7 ± 0.3	0.3 ± 0.3	0.8 ± 0.5
1.5 N	BayesClean I1 (std.)	98.2 ± 1.1	15.1 ± 1.5	58.6 ± 1.0
1.5 N	BayesClean I1 (raw)	98.2 ± 1.1	0.2 ± 0.3	6.3 ± 0.4
Boundary transfer	Unfiltered	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Boundary transfer	Original residual	0.0 ± 0.0	0.3 ± 0.3	0.8 ± 0.5
Boundary transfer	BayesClean I1 (std.)	92.3 ± 5.0	51.7 ± 4.5	57.2 ± 1.2
Boundary transfer	BayesClean I1 (raw)	96.3 ± 2.0	1.2 ± 0.4	14.0 ± 1.7

Table 31: Literal-predicate diagnostics: candidate counts, mean ± sample SD. The two coordinate rules share one posterior per seed and world. A reversal means $|\mu + 0.5\sigma| > |\mu + 1.5\sigma|$; overlap means membership in both printed $\mathcal{I}_1$ and $\mathcal{I}_3$. Only $\mathcal{I}_1$ controls retention.

Label world	Predicate coordinates	$\mathcal{I}_1$ kept	Reversals	$\mathcal{I}_1 \cap \mathcal{I}_3$
All correct	Standardized	2746.4 ± 28.1	2925.4 ± 37.4	627.4 ± 21.6
All correct	Raw newtons	4648.6 ± 35.2	0.2 ± 0.4	0.0 ± 0.0
1.5 N	Standardized	2528.2 ± 59.1	1481.2 ± 20.1	111.6 ± 26.0
1.5 N	Raw newtons	5433.4 ± 41.4	0.0 ± 0.0	0.0 ± 0.0
Boundary transfer	Standardized	2511.2 ± 73.9	2516.0 ± 85.3	327.4 ± 63.8
Boundary transfer	Raw newtons	5014.4 ± 75.2	0.0 ± 0.0	0.0 ± 0.0

Table 32: Rejection within the same supplied-label sign groups, defined by $v = (y - \bar{y})/s_y < 0$ and $v \geq 0$ for both policies. Percentages, mean ± sample SD. These diagnostics do not select a policy. Empty groups receive no rate.

Label world	Predicate coordinates	Rejected, $v < 0$ (%)	Rejected, $v \geq 0$ (%)
All correct	Standardized	75.0 ± 0.9	27.4 ± 0.3
All correct	Raw newtons	20.1 ± 0.6	28.1 ± 0.3
1.5 N	Standardized	79.6 ± 1.0	22.2 ± 0.6
1.5 N	Raw newtons	3.6 ± 0.8	22.5 ± 0.7
Boundary transfer	Standardized	82.0 ± 1.3	26.7 ± 0.9
Boundary transfer	Raw newtons	10.1 ± 1.9	27.6 ± 1.0

Table 33: All per-seed final-head outcomes: All correct. MAEs and within-method sharp attack-minus-correct changes are in N. Every row includes the same 362 verified contacts. Splits reuse one acquisition population.

Seed	Candidate treatment	Ordinary MAE	Sharp MAE	Sharp Δ
0	Unfiltered	0.0275	0.0371	+0.0000
0	Original residual	0.0251	0.0308	+0.0000
0	BayesClean I1 (std.)	0.0263	0.0333	+0.0000
0	BayesClean I1 (raw)	0.0250	0.0342	+0.0000
1	Unfiltered	0.0242	0.0286	+0.0000
1	Original residual	0.0245	0.0312	+0.0000
1	BayesClean I1 (std.)	0.0246	0.0318	+0.0000
1	BayesClean I1 (raw)	0.0245	0.0281	+0.0000
2	Unfiltered	0.0271	0.0455	+0.0000
2	Original residual	0.0305	0.0484	+0.0000
2	BayesClean I1 (std.)	0.0308	0.0494	+0.0000
2	BayesClean I1 (raw)	0.0275	0.0467	+0.0000
3	Unfiltered	0.0273	0.0355	+0.0000
3	Original residual	0.0259	0.0343	+0.0000
3	BayesClean I1 (std.)	0.0268	0.0371	+0.0000
3	BayesClean I1 (raw)	0.0258	0.0358	+0.0000
4	Unfiltered	0.0241	0.0335	+0.0000
4	Original residual	0.0244	0.0349	+0.0000
4	BayesClean I1 (std.)	0.0247	0.0365	+0.0000
4	BayesClean I1 (raw)	0.0242	0.0332	+0.0000

Table 34: All per-seed final-head outcomes: 1.5 N. MAEs and within-method sharp attack-minus-correct changes are in N. Every row includes the same 362 verified contacts. Splits reuse one acquisition population.

Seed	Candidate treatment	Ordinary MAE	Sharp MAE	Sharp Δ
0	Unfiltered	0.0407	0.2330	+0.1959
0	Original residual	0.0266	0.0391	+0.0083
0	BayesClean I1 (std.)	0.0264	0.0381	+0.0048
0	BayesClean I1 (raw)	0.0253	0.0340	-0.0003
1	Unfiltered	0.0335	0.1627	+0.1341
1	Original residual	0.0240	0.0326	+0.0014
1	BayesClean I1 (std.)	0.0263	0.0355	+0.0037
1	BayesClean I1 (raw)	0.0266	0.0366	+0.0085
2	Unfiltered	0.0394	0.1794	+0.1339
2	Original residual	0.0295	0.0484	+0.0000
2	BayesClean I1 (std.)	0.0300	0.0482	-0.0013
2	BayesClean I1 (raw)	0.0304	0.0491	+0.0024
3	Unfiltered	0.0345	0.2245	+0.1890
3	Original residual	0.0249	0.0384	+0.0041
3	BayesClean I1 (std.)	0.0271	0.0394	+0.0023
3	BayesClean I1 (raw)	0.0256	0.0394	+0.0036
4	Unfiltered	0.0342	0.2011	+0.1676
4	Original residual	0.0246	0.0378	+0.0029
4	BayesClean I1 (std.)	0.0259	0.0374	+0.0009
4	BayesClean I1 (raw)	0.0249	0.0341	+0.0009

Table 35: All per-seed final-head outcomes: Boundary transfer. MAEs and within-method sharp attack-minus-correct changes are in N. Every row includes the same 362 verified contacts. Splits reuse one acquisition population.

Seed	Candidate treatment	Ordinary MAE	Sharp MAE	Sharp Δ
0	Unfiltered	0.0346	0.1212	+0.0841
0	Original residual	0.0337	0.1198	+0.0890
0	BayesClean I1 (std.)	0.0299	0.0415	+0.0082
0	BayesClean I1 (raw)	0.0260	0.0364	+0.0022
1	Unfiltered	0.0296	0.0847	+0.0561
1	Original residual	0.0296	0.0991	+0.0679
1	BayesClean I1 (std.)	0.0268	0.0379	+0.0061
1	BayesClean I1 (raw)	0.0237	0.0338	+0.0057
2	Unfiltered	0.0348	0.0982	+0.0526
2	Original residual	0.0362	0.1056	+0.0572
2	BayesClean I1 (std.)	0.0334	0.0536	+0.0041
2	BayesClean I1 (raw)	0.0295	0.0481	+0.0014
3	Unfiltered	0.0301	0.0875	+0.0520
3	Original residual	0.0302	0.0867	+0.0524
3	BayesClean I1 (std.)	0.0270	0.0538	+0.0167
3	BayesClean I1 (raw)	0.0254	0.0402	+0.0044
4	Unfiltered	0.0278	0.0845	+0.0510
4	Original residual	0.0266	0.1043	+0.0693
4	BayesClean I1 (std.)	0.0267	0.0450	+0.0085
4	BayesClean I1 (raw)	0.0243	0.0383	+0.0052

L Budgeted Verified Normalization

The source BayesClean experiments normalize using clean training statistics (Carnerero-Cano et al., 2026), whereas Appendix K uses supplied candidates. We therefore freeze a further sensitivity comparison before fitting: replace both feature and target means and population standard deviations by moments over the rows of the same 362 verified contacts. This normalizer is identical across the three label worlds. The Bayesian likelihood still contains only supplied candidate rows, not verified rows. The final learner, as before, retains every verified row. Thus the total available label budget is unchanged, but the sanitizer now accesses those verified labels for normalization. Both transforms change together; this does not isolate target centering alone.

All candidate indices, supplied labels, test contacts, solver settings, two I1 coordinates, initializations and final 2,000-update schedules remain fixed. This adds 30 neural fits and 15 shared Bayesian fits over the same five exploratory splits. Using an available verified subset is not the same as observing the complete pre-poison clean training population in the source benchmark. The results remain explicitly labelled adaptations, not an author-code reproduction. The unchanged old boundary labels remain a transfer test rather than a sanitizer-adaptive attack.

For 1.5 N replacements, the standardized/raw variants change 18.47%/10.85% of candidate decisions relative to supplied normalization. The standardized rule's ordinary loss decreases from 58.59% to 49.39%, whereas the raw rule's increases from 6.29% to 15.61%; both reach 98.95% changed-label recall. Their paired mean sharp-MAE changes are $+0.0003 \pm 0.0014$ and $+0.0009 \pm 0.0042$ N. These five-split dispersions do not establish equivalence or a uniform utility improvement. Under frozen boundary labels, raw I1 recall rises from 96.32% to 98.36%, but correct-cover loss rises from 1.18% to 4.27%; its sharp MAE changes from 0.0394 to 0.0398 N. The own-clean paired excess decreases from 0.0038 to 0.0024 N while clean MAE itself increases from 0.0356 to 0.0373 N. Reporting the clean reference prevents interpreting that smaller excess alone as better final prediction.

The literal I1 predicate can reject a correctly predicted negative target even with a clean normalizer. For $\mu = y$, $c > 0$ and $\sigma > 0$,

$$|y| \le |y + c\sigma| \quad \Longleftrightarrow \quad y \ge -c\sigma/2.$$

This elementary identity concerns the predicate at a chosen origin, not a claim that the fitted predictor has zero error. Clean normalization does not turn I1 into a symmetric residual band or eliminate its coordinate dependence. The tables report both fixed coordinates without selecting one from observed outcomes. Deletion statistics concern candidates; predictive utility also benefits from the fully retained verified data.

A separate NumPy/SciPy calculation reconstructs verified moments, posterior linear solves and evidence updates. It checks both predicates, all retained inputs and labels, scalers, initialization and minibatch records, and independently reloads all 30 final neural states to reproduce the saved predictions. Ordinary/sharp trajectory MAE, paired contrasts and filtering metrics are recalculated from records. This is an independent code path, not external replication or a replay of SGD optimization. The supplement retains all seed-level results and source hashes.

Table 36: Verified-normalization sensitivity (five-seed mean ± SD, N). Every model retains all 362 verified contacts. Δ_{clean} compares the same coordinate with its own all-correct world; Δ_{cand} compares the matching candidate-normalized pipeline in Table 3. Both are paired end-to-end differences, not fixed-pool label effects.

Labels	Rule	Ordinary MAE	Sharp MAE	Δ_{clean}	Δ_{cand}
All correct	Std.	0.0272 ± 0.0018	0.0392 ± 0.0061	0.0000 ± 0.0000	0.0015 ± 0.0012
All correct	Raw	0.0260 ± 0.0016	0.0373 ± 0.0071	0.0000 ± 0.0000	0.0017 ± 0.0020
1.5 N	Std.	0.0278 ± 0.0016	0.0400 ± 0.0049	0.0009 ± 0.0018	0.0003 ± 0.0014
1.5 N	Raw	0.0271 ± 0.0021	0.0395 ± 0.0060	0.0022 ± 0.0022	0.0009 ± 0.0042
Boundary transfer	Std.	0.0276 ± 0.0021	0.0461 ± 0.0069	0.0070 ± 0.0042	-0.0002 ± 0.0021
Boundary transfer	Raw	0.0262 ± 0.0013	0.0398 ± 0.0053	0.0024 ± 0.0018	0.0004 ± 0.0015

Table 37: Verified-normalized I1 filtering (percentages; five-seed mean ± SD). Recall concerns actually changed labels and is undefined in the all-correct world. Cover and ordinary loss use their respective correct candidate denominators. Disagreement is the fraction of all candidate decisions differing from the corresponding candidate-normalized rule. Verified rows are never deleted.

Labels	Rule	Recall (%)	Cover loss (%)	Ord. loss (%)	Disagree (%)
All correct	Std.	–	60.20 ± 3.38	54.27 ± 0.71	2.21 ± 0.69
All correct	Raw	–	30.46 ± 2.85	23.20 ± 0.67	1.43 ± 0.78
1.5 N	Std.	98.95 ± 0.95	17.37 ± 1.84	49.39 ± 2.24	18.47 ± 1.50
1.5 N	Raw	98.95 ± 0.95	2.70 ± 1.15	15.61 ± 0.71	10.85 ± 0.77
Boundary transfer	Std.	93.96 ± 3.91	52.30 ± 4.03	53.60 ± 1.20	12.17 ± 0.95
Boundary transfer	Raw	98.36 ± 0.92	4.27 ± 1.64	20.17 ± 1.27	9.31 ± 0.76

M SANITIZER-AWARE I1 STRESS TEST

Appendix L changes normalization but transfers labels built for a different residual screen. To test the fitted sanitizer directly, we lock a label-only attack before all fits. This follows inspection of the preceding experiments and reuses their five test partitions, so it is exploratory rather than an independent replication.

Knowledge and construction. The attacker knows candidate features and supplied labels, the 362-contact verified normalizer, posterior algorithm, and raw-unit I1 predicate. It cannot alter features, verified labels, source assignments, cover labels, or any test quantity. Let P be the same poison-position mask as in the 1.5 N study. Starting with those 1.5 N replacements, iteration k refits the posterior and sets, only for $i \in P$,

$$y_i^{(k+1)} = \text{float32}\Big[\text{clip}\left(\left|\mu_i^{(k)} + \tfrac{1}{2}\sigma_i^{(k)}\right| - 10^{-4}, 0, 1.5\right)\Big]. \tag{31}$$

Exactly 20 query/update rounds are used. A distinct 21st posterior fit applies the actual final I1 mask; no iterate is selected, repaired, or early-stopped. This is a target-seeking fixed-point heuristic, not bilevel victim-loss optimization. It has stronger sanitizer knowledge than the fixed attack but a different label magnitude, so recall values are not equal-strength comparisons.

For every seed, three native heads use the same initialization and 2,000-update protocol: adaptive labels after final filtering; the *same retained inputs* with correct labels restored; and adaptive labels without filtering. The first pair has identical input scaler and complete minibatch sequence. The restored-label arm is an oracle estimand for label effect, not a deployable defense. All verified rows remain in every model.

The filter rejects 29.16 ± 11.36 percentage points of changed positions and retains an average of 215.4 corrupted rows. It also removes $25.13 \pm 2.61\%$ of correct sharp cover and $24.44 \pm 0.50\%$ of ordinary candidates. Adaptive labels change by $+0.05470 \pm 0.01046$ N from truth on average; $24.08 \pm 4.66\%$ move downward, despite the target-seeking update. The mean shift among retained poison positions is $+0.04488 \pm 0.00802$ N. Relative to the raw-I1 all-correct pipeline, filtered sharp MAE increases by 0.00322 ± 0.00249 N; this full-pipeline contrast includes a different retained set and is distinct from the same-input effect in Table 38.

Table 38: Sanitizer-aware stress test, mean ± sample SD over five reused splits (N). The paired effect holds retained inputs fixed.

Training arm	Ordinary MAE	Sharp MAE	Sharp signed bias
Adaptive, filtered	0.02712 ± 0.00234	0.04053 ± 0.00812	0.00683 ± 0.00619
Same inputs, correct labels	0.02694 ± 0.00202	0.03568 ± 0.00738	0.00301 ± 0.00376
Adaptive, unfiltered	0.02609 ± 0.00157	0.04351 ± 0.00708	0.01350 ± 0.00706
Paired filtered label effect	0.00019 ± 0.00066	0.00485 ± 0.00170	–

Table 39: Every prescribed split. "Rejected" is evaluated by the fresh final posterior; Δ_{sharp} is adaptive-filtered minus matched-correct MAE.

Seed	Rejected poison positions	Last max update (N)	Δ_{sharp} (N)
0	103/304	0.002686	0.004221
1	136/302	0.004969	0.003281
2	60/303	0.000714	0.006156
3	93/307	0.004145	0.003477
4	51/304	0.000543	0.007113

Negative result and audit. Final acceptance is incomplete in every seed, and the last update remains nonzero (Table 39). Refitting the posterior moves its boundary; placement inside the preceding boundary is not a fixed-point guarantee. Thus the experiment closes the narrower transferred-attack gap, but neither proves convergence nor evaluates a worst-case adaptive attack, standardized I1, different query budgets, or independent acquisitions.

The run contains 100 attacker posterior queries, five final defender fits, and 15 neural fits. An estimator-free NumPy/SciPy path reconstructs all 105 posterior solutions, evidence updates, label updates, and masks. It reconstructs omitted intermediate covariance matrices from saved precisions, checks all five saved final covariances, and independently reloads all 15 neural states for exact prediction equality. The audit passes 34,001 assertions. It does not replay SGD, establish attack optimality, or constitute external replication. All query outputs, states, sampling records, predictions, and source hashes are retained.

N Execution Records and Reproduction

The artifact retains pinned provenance, complete contact assignments, selected observations, true and supplied labels, predictions, filtering decisions, reference parameters and neural training records. Model states and feature banks are saved separately from the compact supplement. The README gives upstream acquisition and distinguishes record exports from fitting and audits.

Original protocols preceded their held-out predictions; later studies explicitly reuse that finite population. Eligibility, batch budgets and matched controls were fixed before formal execution. No favorable seed was selected. Synthetic integration and runtime checks are excluded from tactile results. Repeated fits do not estimate device, gel or laboratory variation.

Execution counts and measured computation. The original studies perform 225 neural fits. Sparsh feature extraction takes 2,918.0 s; the attack, screening, mitigation and batch-shift programs take 1,222.1, 22.0, 646.3 and 593.6 s, respectively. These CPU wall times overlap and must not be summed as elapsed time. The second revision adds 25 neural fits and 20 final plus 80 inner Ridge fits. The third adds 105 neural fits, including 70 second-encoder heads; DINO extraction takes 974.0 s. Its curation reference uses five final and 20 inner Ridge fits. A separate 25-fit float64 DINO-reference diagnostic leaves its fixed-target curation decisions unchanged.

The total-budget study adds 60 neural fitting calls, plus 15 final and 60 inner Ridge fits. The count-derived stopping sensitivity adds 15 neural fits, bringing that stage's formal total to 430. The category-exclusion follow-up adds 30 further heads, bringing that stage to 460 fitting calls. The label-aware sanitizer follow-up adds 50 neural fits and 15 shared Bayesian fits; ten comparison cells

reuse unchanged earlier neural states. Its new per-fit wall times sum to 949.2 s for neural training and 19.4 s for Bayesian fits, with two concurrent single-thread workers; these sums are not elapsed runtime. That stage contains 510 neural fitting calls. The verified-normalization sensitivity adds 30 neural fits and 15 shared Bayesian fits, bringing that stage to 540 neural fitting calls. The sanitizer-aware stress test adds 15 neural fits, 100 attacker posterior queries and five final defender posterior fits, bringing the complete study to 555 neural fitting calls, including five deliberate replays of earlier full-budget controls. Copies of reused states and tiny synthetic software checks add no tactile fitting calls or independent acquisition evidence. All extensions retain complete records and separate audits. The total-budget audit reconstructs reference and curation calculations; the exposure audit reloads all 15 shorter-trained heads and checks their original minibatch prefixes. The category-exclusion audit reconstructs both added pools, reloads all 30 new heads and recomputes earlier comparison predictions. None of these audits fits another model. Exact collateral frontiers and calibration-budget thinning require no neural fitting. Package versions, commands, source hashes and per-fit times are retained.

Interrupted execution and recovery. An incomplete eight-thread pilot was excluded before the original formal runs. Later deterministic-record restorations reproduced recoverable arrays and prior sampling/calibration hashes without altering fitted models or predictions. A DINO worker-reuse setup error occurred after 42 heads; only the remaining 28 prescribed heads were subsequently executed, preserving completed-file hashes. The total-budget startup failed during optimizer construction before any neural update; its five reference and 20 inner Ridge fits are excluded from formal counts. Missing dependencies were installed before executing all prescribed neural fits. Original files, logs, dispositions and reconstruction records remain available. These events change neither the reported scientific settings nor the retained seed set.