Auditing Offline Demonstration Pruning for Online Robot Deployment

Anonymous Authors[1]

Abstract

Pruning offline demonstrations is a deployment decision: a scoring rule can beat random pruning while still losing to a cheap baseline that would be the better online choice. We present a four-gate offline-to-online audit for pruning methods: paired random comparison, cheap-baseline audit, held-out stress test, and mechanism check. As a case study, TRAK-Traj, a trajectory-level adaptation of TRAK-style attribution, beats matched random pruning on curated RoboMimic Can MH by +4.7 percentage points across 10 paired seeds and 300 rollouts per condition (9/10 wins; paired $t(9) = 2.43$, one-sided $p = 0.019$). But the audit changes the deployment recommendation: TracIn-style scoring ties TRAK, trajectory length is stronger on curated Can, and a held-out mixed-quality block shows length outperforming both TRAK and random. Mechanism analysis explains why: length pruning removes 260 worse-tier trajectories out of 270 pruned in the mixed-quality split. The contribution is a reproducible audit template for data-curation claims before robot deployment.

1. Introduction

Offline robot learning is often limited less by the amount of data than by the quality and consistency of the demonstrations available for training. Behavior cloning and imitation learning can turn logged demonstrations into deployable policies (Pomerleau, 1989; Ross et al., 2011; Osa et al., 2018), but multi-human datasets can contain strong demonstrations, hesitant executions, corrective behavior, and failed strategies (Mandlekar et al., 2021). Training on all trajectories is the default, but it is not a neutral choice: the learner may average across inconsistent action modes or spend limited optimization budget fitting trajectories that are not useful for deployment.

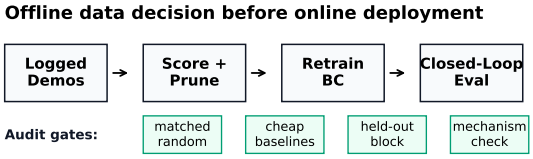

Figure 1. Offline-to-online audit view. The intervention is an offline data decision; the outcome is closed-loop success after retraining. We require matched random, cheap-baseline, held-out, and mechanism checks before turning a pruning result into a deployment recommendation.

This paper asks how to evaluate the offline decision to prune demonstrations under a fixed retraining budget. This fits the offline-to-online workflow studied in offline RL and offline imitation learning: a batch dataset is processed before the learned policy is evaluated in closed-loop deployment (Fu et al., 2020; Levine et al., 2020). We study the simplest operational form: assign one score to each trajectory, remove the lowest-scoring trajectories, retrain behavior cloning from scratch on the retained set, and ask whether the resulting online policy improves. TRAK-Traj is our attribution case study, but the audit is the main contribution: we propose a four-gate protocol that checks matched random pruning, cheap deterministic baselines, held-out mixed-quality seeds, and mechanism evidence before turning a pruning result into a deployment recommendation.

We propose TRAK-Traj, a trajectory-level adaptation of TRAK (Park et al., 2023). The method replaces per-example attribution with per-trajectory attribution, the natural unit at which robot demonstrations are collected and filtered. The goal is not to solve online adaptation directly, but to improve the offline data substrate from which policies are learned.

The key design choice is that the evaluation is cold-start and paired. After pruning, every policy is re-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

trained from random initialization with the same architecture and budget. Random pruning and TRAK-Traj pruning share the same seed and prune ratio, so the comparison asks whether the ranking itself adds value. This is a demanding test for a 10% pruning method: the method removes only 9 of 90 curated demonstrations, leaving little room for large effects.

Our primary evidence is a high-rollout cold-start evaluation on RoboMimic Can MH (Mandlekar et al., 2021). On curated demonstrations, TRAK-Traj improves over matched random pruning in 9 of 10 seeds. The average improvement is +4.7 percentage points, with complementary parametric, non-parametric, permutation, bootstrap, and Bayesian checks supporting a positive effect. A 10-seed deterministic-baseline audit changes the practical recommendation: loss is slightly below TRAK, TracIn-style scoring ties TRAK on mean success, and trajectory length is higher. On a held-out full mixed-quality block, TRAK does not beat random on mean success (.143 vs. .162), while length is substantially stronger (.278). This motivates a sharper conclusion: attribution is a useful candidate signal in the controlled curated protocol, but random-only evaluation would overstate its practical value.

Our contributions are:

- A four-gate cold-start audit protocol for offline robot demonstration pruning.
- A trajectory-level TRAK case study showing a statistically significant advantage over matched random pruning.
- A completed cheap-baseline audit showing that this positive random-baseline result does not imply practical dominance.
- A mechanism analysis explaining why trajectory length is strong in Can MH and why attribution and length prune largely different demonstrations.

2. Related Work

Offline robot imitation and decision making. Behavior cloning from human demonstrations remains a common foundation for robot manipulation (Pomerleau, 1989; Ross et al., 2011; Osa et al., 2018), but performance is sensitive to demonstration source and quality. RoboMimic provides standardized multi-human datasets and evaluation protocols for studying these effects (Mandlekar et al., 2021), while modern robot policy classes such as diffusion policies and action-chunking transformers further increase the importance of the offline training distribution (Chi et al., 2023; Zhao et al., 2023). Offline RL benchmarks and algorithms similarly emphasize that batch data quality and support determine whether a policy can be safely improved before deployment (Fu et al., 2020; Kumar et al., 2020; Kostrikov et al., 2022). Recent imitation-learning work formalizes data quality as a central object rather than a nuisance variable (Belkhale et al., 2023). Data generation and curation systems such as MimicGen (Mandlekar et al., 2023), IntervenGen (Hoque et al., 2024), and ReMix (Hejna et al., 2025) show that policy quality depends on which demonstrations are included, not merely on dataset size.

Training data attribution. Influence functions (Koh & Liang, 2017), TracIn (Pruthi et al., 2020), datamodels (Ilyas et al., 2022), and TRAK (Park et al., 2023) estimate how training data affect model behavior. TRAK scales attribution through projected gradients and linearized predictors. TRAK-Traj follows this line but changes the attribution unit from examples to complete trajectories, aligning the method with robot data collection and pruning. This lineage also motivates caution: influence-style linearizations can be fragile in deep nonconvex models or answer a proxy question rather than exact leave-one-out retraining (Basu et al., 2021; Bae et al., 2022).

Data valuation and pruning. Data Shapley (Ghorbani & Zou, 2019), DVRL (Yoon et al., 2020), coreset-style selection (Sener & Savarese, 2018), forgetting events (Toneva et al., 2019), dataset cartography (Swayamdipta et al., 2020), and gradient-matching subset selection (Killamsetty et al., 2021) study how to value or select training data. Our work is closest in spirit to practical pruning: score once, remove low-value units, and retrain under the same budget. The audit result is aligned with recent data-centric lessons: simple dataset statistics can be strong baselines and should be reported before claiming that a learned valuation method is practically dominant.

3. Method

3.1. Problem Setup

Let $\mathcal{D} = \{\tau_i\}_{i=1}^{N}$ be an offline imitation dataset, where each trajectory is

$$\tau_i = \{(o_t, a_t)\}_{t=1}^{T_i}. \tag{1}$$

We train a behavior cloning policy f_θ and seek a scalar score s_i for each full trajectory. For prune ratio p, we retain the top $(1-p)N$ trajectories by score and retrain from random initialization using unchanged architecture, optimizer, and training budget. This cold-start

design makes the evaluation conservative: gains must come from the retained data, not from warm-starting or per-condition tuning.

3.2. Trajectory-Level TRAK

TRAK-Traj adapts TRAK-style attribution to the demonstration level. For checkpoint c, we define a trajectory surrogate objective $q_{\theta^{(c)}}(\tau_i)$ and compute its parameter gradient

$$g_i^{(c)} = \nabla_\theta q_{\theta^{(c)}}(\tau_i). \tag{2}$$

In our experiments, q is computed over sampled transitions from the trajectory, and gradients are taken with respect to the policy head for efficiency. For behavior cloning, the surrogate is label-agnostic:

$$q_\theta(\tau_i) = \frac{1}{|\mathcal{S}_i|} \sum_{(o,a)\in\mathcal{S}_i} \|f_\theta(o)\|_2^2, \tag{3}$$

where $\mathcal{S}_i$ is a deterministic sample of 16 transitions from trajectory τ_i. This follows the D-TRAK-style intuition that gradients of model output energy can provide a useful local linearization without using the action label directly in the attribution objective (Zheng et al., 2024). The policy is still trained with standard behavior cloning loss; the surrogate is used only for scoring.

We project gradients using a Johnson–Lindenstrauss matrix $R \in \mathbb{R}^{d\times P}$:

$$\phi_i^{(c)} = R g_i^{(c)}, \quad d = 2048. \tag{4}$$

For each checkpoint, projected features are ridge-preconditioned:

$$H^{(c)} = \frac{1}{N} \sum_i \phi_i^{(c)} \phi_i^{(c)\top} + \lambda I. \tag{5}$$

Given target features from the held-out better_valid split, the score is

$$s_i = \frac{1}{C} \sum_{c=1}^{C} \phi_i^{(c)\top} (H^{(c)})^{-1} \psi^{(c)}. \tag{6}$$

Operationally, the implementation featurizes the exact training trajectories used by the baseline and the held-out target trajectories, builds the ridge matrix from the training features, solves for the target direction, and scores each training trajectory by its dot product with that direction. Higher scores are kept; lower scores are pruned.

We average ranks across three seed offsets $\{0, 97, 211\}$ and multiple saved checkpoints to reduce variance. Before aggregation, member score vectors are winsorized at ± 2 standard deviations, then converted to ranks and averaged. This makes the pruning decision depend on relative order rather than raw score scale, which varies across checkpoints and projection seeds.

Why target high-quality validation trajectories? The target set is not used for retraining. It defines the behavior whose linearized features the method should preserve. In Can MH, better_valid provides a small held-out set from the same operator-quality tier as the curated training split. In the mixed-quality experiment, using the same target split tests whether a high-quality target direction can identify useful demonstrations inside the full heterogeneous training set. This is a practical but not assumption-free design: it presumes access to a held-out validation subset whose quality is representative of the desired behavior.

3.3. Pruning Pipeline

The full pipeline is:

1. Train a baseline behavior cloning policy on the available offline dataset.
2. Score each trajectory with TRAK-Traj.
3. Remove the bottom p fraction of trajectories.
4. Retrain the policy from scratch on the retained subset.
5. Evaluate closed-loop success with strict success parsing.

4. Experimental Setup

Task. We evaluate on RoboMimic Can MH (PickPlaceCan), a multi-human manipulation dataset with 300 demonstrations across operator quality tiers. We use the standard 90/10 train-validation split. The curated condition uses the 90 "better" training demonstrations. The mixed-quality condition uses all 270 training demonstrations.

Why lead with curated data? The curated condition is not the most realistic deployment setting, but it is the cleanest controlled test of whether attribution improves over random removal when every demonstration is nominally useful. Because only 9 of 90 trajectories are removed, random pruning is already a strong baseline and the expected effect is small. The mixed-quality condition is closer to the motivating data-curation use case, but its fixed-budget behavior

cloning baselines can collapse. We therefore treat curated Can as the confirmatory comparison and mixed-quality Can as a stress test.

Policy. All conditions use a residual MLP behavior cloning policy with 5 residual layers, hidden dimension 768, dropout 0.05, observation history 2, and action history 1. Training uses Adam with learning rate 3×10^{-4}, batch size 128, and 8000 gradient updates.

Attribution configuration. TRAK-Traj uses projection dimension 2048, 16 trajectory samples for gradient estimation, fp32 precision, three split seed offsets $\{0, 97, 211\}$, rank-mean aggregation, head-only gradients, and score winsorization at $\pm 2\sigma$.

Evaluation. Each reported condition is evaluated with 300 environment rollouts using state-anchored reset, per-rollout seeds, and strict success parsing. This follows recent calls for robot-learning evaluations to report rollout counts, success criteria, and statistical analysis rather than only aggregate success rates (Kress-Gazit et al., 2024). In the curated condition, we run 10 paired seeds (607–616). In the mixed-quality condition, we use seeds 701–710 as an exploratory scan and seeds 711–716 as a held-out audit block. We treat mixed-quality results as diagnostic because several seeds are near the floor and because the mixed-quality setting has higher seed-to-seed variance.

Baselines. The main confirmatory baseline is matched random pruning: for each seed and prune ratio, a random subset of the same size is removed and the policy is retrained from scratch. This directly tests whether attribution provides information beyond deleting the same amount of data. We also audit cheap deterministic comparators, including loss, trajectory length, and TracIn-style scores. That audit is diagnostic rather than primary: it tests whether a positive result against random survives contact with obvious low-cost alternatives.

5. Results

5.1. Curated Demonstrations: TRAK Beats Random

Figure 2 and Table 1 show the primary result. TRAK-Traj improves over random pruning in 9 of 10 seeds. The mean gain over random is +4.7 percentage points, despite pruning only 9 of 90 trajectories. TRAK also exceeds the unpruned baseline mean and is higher than the unpruned baseline in 6 of 10 seeds. We interpret this as evidence that the ranking can remove trajectories that are harmful or redundant for this training regime; it is not, by itself, evidence that the same score will dominate all pruning heuristics or all robot tasks.

Table 1. Can MH curated demonstrations at 10% prune ratio. Each seed uses 300 rollouts per condition.

Seed	Baseline	Random	TRAK	Δ
607	.863	.770	.800	+.030
608	.723	.780	.823	+.043
609	.617	.797	.827	+.030
610	.793	.680	.803	+.123
611	.800	.770	.870	+.100
612	.783	.727	.863	+.137
613	.787	.767	.783	+.017
614	.837	.847	.773	-.073
615	.770	.773	.830	+.057
616	.810	.787	.797	+.010
Mean	.778	.770	.817	+.047

Table 2. Statistical checks for the curated Can MH comparison. Tests are one-sided for the directional hypothesis TRAK > random.

Check	Statistic	Result
Paired t-test	$t(9) = 2.43$	$p = 0.019$
Wilcoxon	p_{1s}	0.019
Sign test	9/10 wins	$p = 0.011$
Permutation	50K draws	$p = 0.023$
Bootstrap CI	10K resamples	$[+1.0, +8.3]$ pp
Cohen's d_z	paired effect	0.77
Bayesian	$P(\mu > 0)$	0.993

Table 2 summarizes complementary statistical checks. The conclusion is not driven by one test family: parametric, non-parametric, permutation, bootstrap, and Bayesian summaries (Benavoli et al., 2017) all support a positive TRAK-Traj advantage in this paired setting. We treat these as robustness checks on one paired claim rather than separate discoveries; therefore we do not apply a multiple-testing procedure such as Benjamini–Hochberg false-discovery-rate control (Benjamini & Hochberg, 1995). The effect is modest in absolute size but meaningful for a minimal intervention: the training set changes by only 10%, and every pruned policy is retrained from scratch. Power calculations help keep this scope realistic (Cohen, 1988).

5.2. Cheap-Heuristic Audit

Table 3 is the current check against simple non-random comparators. It does not support a broad claim that TRAK-Traj dominates every cheap heuristic: loss is slightly below TRAK on mean success (.802 vs. .817), TracIn-style scoring ties TRAK (.817), and trajectory length is higher (.845). This strengthens the need for the scoped claim used here. The established result is that TRAK-Traj provides a reproducible attribution signal beyond matched random pruning in the curated paired protocol; practical dominance over cheaper scor-

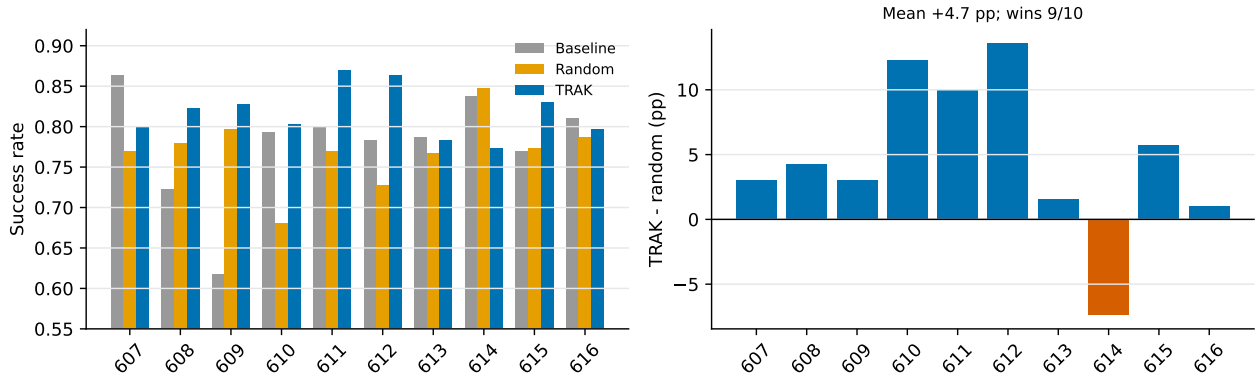

Figure 2. Primary Can MH curated-data result. Left: per-seed closed-loop success rates for the unpruned baseline, matched random pruning, and TRAK-Traj pruning at 10%. Right: paired seed-level deltas show that TRAK-Traj beats random pruning in 9 of 10 seeds, with a +4.7 percentage point mean advantage.

Table 3. Curated Can MH cheap-heuristic audit at 10% pruning. Each entry uses 300 rollouts. This table is an audit, not the primary confirmatory comparison.

Seed	TRAK	Loss	TracIn	Length
607	.800	.777	.870	.857
608	.823	.750	.827	.873
609	.827	.773	.833	.847
610	.803	.857	.827	.897
611	.870	.887	.867	.890
612	.863	.800	.770	.803
613	.783	.703	.777	.813
614	.773	.823	.753	.777
615	.830	.793	.867	.827
616	.797	.860	.780	.870
Completed seeds	10	10	10	10
Mean available	.817	.802	.817	.845

Table 4. Mechanism audit for 10% pruning. Length pruning removes much longer trajectories than TRAK and, in the mixed-quality split, almost directly targets the worse tier. "Overlap" is the mean intersection between TRAK-pruned and length-pruned sets.

Split	TRAK len.	Rule len.	Overlap	Worse
Curated	139.3	196.9	0.7/9	–
Mixed	208.2	462.5	1.9/27	260/270

ing rules is not established by these experiments.

5.3. Why the Length Baseline Wins

Table 4 explains why the cheap baseline is not a nuisance detail. In curated Can, all training demonstrations are from the better tier, yet length pruning removes the longest better demonstrations and achieves the best mean success. In mixed-quality Can, the mechanism is starker: across 10 seeds, length pruning removes 260 worse-tier trajectories out of 270 pruned, while TRAK removes a mixed set of 68 better, 107 okay, and 95 worse trajectories. The pruned-set overlap is small, so length is not merely approximating TRAK; it is exploiting a strong duration-quality confound in this benchmark.

Table 5. Curated Can MH sensitivity across prune ratios. The 10% ratio is the primary minimal-pruning setting; 20% and 30% are exploratory.

Prune	Random	TRAK	Δ	Wins
10%	.770	.817	+.047	9/10
20%	.774	.751	-.022	3/10
30%	.749	.777	+.028	5/10

Table 6. Held-out mixed-quality Can MH audit, seeds 711–716, at 10% pruning. Each condition uses 300 rollouts.

Seed	Baseline	Random	Length	TRAK
711	.163	.100	.263	.127
712	.023	.200	.257	.020
713	.017	.027	.250	.333
714	.303	.387	.253	.167
715	.267	.203	.353	.103
716	.100	.057	.290	.107
Mean	.146	.162	.278	.143
Wins vs. random	–	–	5/6	3/6

5.4. Prune-Ratio Sensitivity

Table 5 shows that the result should not be interpreted as a monotonic pruning law. TRAK-Traj is strongest in the minimal 10% deletion regime and remains directionally positive at 30%, but underperforms random at 20%. This reinforces the paper's scoped claim: the current evidence supports trajectory attribution as a useful ranking signal under a controlled pruning protocol, not as an automatically optimal prune-ratio selector.

5.5. Mixed-Quality Data: Held-Out Audit

Figure 3 and Table 6 show the held-out mixed-quality audit. This block is not favorable to a TRAK-forward

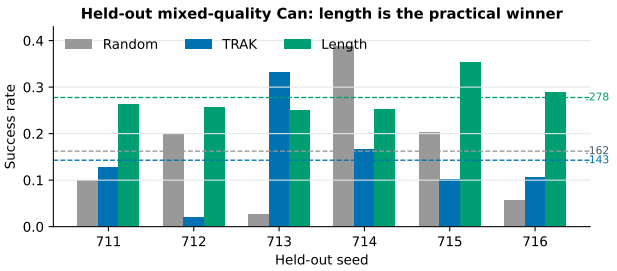

Figure 3. Held-out mixed-quality Can audit. Length is the strongest practical pruning rule in this regime; TRAK does not beat random on mean success. Dashed lines show method means.

Table 7. Scope of the current evidence.

The paper claims	The paper does not claim
TRAK-Traj beats matched random pruning in one 10-seed curated Can MH paired protocol.	TRAK-Traj is a general robot data-curation solution.
Trajectory-level attribution is a plausible offline data-selection primitive for imitation learning.	Attribution always beats cheap heuristics such as length, loss, or TracIn.
The mixed-quality experiment is a useful stress test that exposes a stronger length heuristic.	Mixed-quality Can is a second TRAK confirmatory win.
Minimal pruning can help, but the best prune ratio is empirical.	The score automatically selects the optimal deletion fraction.

method claim: TRAK averages .143 success versus .162 for random and wins only 3 of 6 paired seeds. The length baseline averages .278 and wins 5 of 6 paired seeds against random. The earlier exploratory scan on seeds 701–710 was directionally positive for TRAK (+4.2pp over random, 5/10 wins), but the held-out block shows that this effect is not stable in the full mixed-quality regime. The practical conclusion is therefore the audit conclusion: in heterogeneous Can MH, trajectory duration is the stronger and more reliable pruning signal, while TRAK remains supported only in the curated paired protocol.

5.6. Claim Boundaries

Table 7 states the intended interpretation explicitly. The current evidence is enough for a focused workshop contribution about offline data selection and evaluation, but not enough for a broad deployment claim.

Table 8. Reviewer-facing benchmark card for the current evidence package.

Item	Status
Paired seeds	Yes, 10 curated seeds
Rollouts	300 per condition
Strict success parsing	Yes
Matched random baseline	Yes
Cheap heuristic baselines	Complete 10-seed audit, competitive
Prune-ratio sensitivity	Yes, 10/20/30%
Mixed-quality stress test	Yes, held-out audit supports length
Regeneration command	Yes

5.7. Benchmark Card

Table 8 summarizes the evidence package in the form reviewers can audit quickly. Its most important entry is the completed heuristic audit: the current method has a clear advantage over random pruning in the curated paired protocol, but simple baselines remain serious comparators rather than afterthoughts.

6. Discussion

Why this belongs in offline decision making. TRAK-Traj treats a robot demonstration dataset as logged decision data. The method does not require online interaction during scoring, and retraining evaluates whether an offline subset produces better closed-loop decisions. This makes it a practical bridge between data-centric supervised learning and offline policy learning: the intervention is an offline data decision, while the outcome is online closed-loop success.

What the current evidence supports. The curated Can MH result is a strong focused workshop result: 10 seeds, paired comparisons, 300 rollouts per condition, cold-start retraining, strict success parsing, and consistent wins against matched random pruning. The heuristic audit shows that simple deterministic rules are competitive, with length outperforming TRAK on mean success in this setting. The held-out mixed-quality result strengthens the audit story rather than the TRAK method story: length beats both random and TRAK, while TRAK does not replicate its exploratory mixed-quality gain. We therefore present TRAK-Traj as an empirically supported attribution primitive in the curated protocol, not as a solved general-purpose robot-data filter.

Why not only use full mixed-quality data? The mixed-quality result is closer to the setting practitioners care about, but it is also harder to interpret because the base learner itself is unstable. A failed full-data baseline can mean the dataset contains harmful demonstrations, but it can also mean the fixed training budget, architecture, or optimization seed failed

to find a competent policy. Attribution can only help if the scoring policy and target features contain usable signal. This is why we report both regimes: curated Can establishes a controlled positive result, while mixed-quality Can exposes the deployment risk and the duration-quality confound that motivated the audit.

Limitations. The current confirmatory result is on one task family, one prune ratio, and one behavior cloning architecture. The primary comparison is against matched random pruning; comparisons against cheap deterministic heuristics and learned data valuation methods are necessary before claiming practical dominance. The mixed-quality setting is more realistic but also more unstable. Future work should expand to additional tasks, test diffusion policies, and evaluate scorer robustness when no competent baseline policy is available.

7. How To Use the Audit

The point of the audit is to prevent an offline data decision from becoming a hidden deployment risk. In a normal robot-learning workflow, the pruning method is chosen before online evaluation is available. It is therefore tempting to select the method that beats random in an exploratory run and then report the resulting policy improvement as evidence that the scoring rule is practically useful. Our results show why that is insufficient. TRAK-Traj has a real positive signal in the curated paired protocol, but the held-out mixed-quality block and length baseline change the operational recommendation.

Decision rule. We recommend treating a pruning method as deployment-ready only if it passes four gates. First, it should beat matched random pruning in a paired seed-level comparison. Second, it should remain competitive with cheap deterministic baselines that require no attribution pass. Third, it should be checked on a held-out seed block or task variant that was not used to choose the story. Fourth, the selected and removed trajectories should be inspected for a plausible mechanism. A method that passes the first gate but fails the second or third can still be scientifically interesting, but it should be described as a candidate signal rather than a deployment recommendation.

What changes for Offline2Online. For the Offline2Online workshop, the relevant object is not only the attribution method but the offline decision policy: which data should be kept before the robot is evaluated online? Table 9 makes the current answer explicit. On curated Can, TRAK-Traj is a valid candidate signal because it repeatedly beats matched random pruning under cold-start retraining. On mixed-quality Can, the stronger decision rule is length pruning, not TRAK. This is still a useful offline-to-online result: it identifies a simple data statistic that improves deployment success in a heterogeneous dataset and shows why attribution-only evaluation would have selected the wrong practical conclusion.

Table 9. Deployment interpretation of the audit outcomes.

Evidence	Interpretation	Action
TRAK > random	attribution signal exists	keep studying
Length > TRAK	cheap rule stronger	do not claim dominance
Held-out reversal	mixed regime unstable	scope the claim
Mechanism found	duration-quality confound	report and exploit

Why the negative held-out result helps. The held-out mixed-quality result makes the paper less like a method advertisement and more like an audit of an offline decision. That is the better workshop fit. Offline-to-online learning papers often need to decide what to trust before online interaction is available. A result that only says "our method beats random" gives little guidance when a cheap baseline is stronger. A result that says "the first comparison is positive, but the deployment recommendation changes after cheap-baseline and held-out checks" gives a concrete evaluation recipe for future data-curation systems.

8. Reproducibility Notes

The experiments use the public RoboMimic Can MH low-dimensional dataset and the fixed configuration files configs/pipeline/cold_start_can.yaml and configs/pipeline/full_train_can.yaml. The curated runs use task.demo_split=better_train; mixed-quality runs use task.demo_split=train. Reported success rates are parsed from runs/*_pipeline/metrics.jsonl; the workshop figures and tables are generated from these artifacts rather than manually entered evaluation logs. The paper assets can be regenerated with python3 scripts/build_paper_assets.py.

For the curated result, the source run directories are the can_cold_can_mh_s607–s616 pipelines. For the exploratory mixed-quality scan, the source run directories are can_full_can_mh_s701–s710; for the held-out mixed-quality audit, they are can_full_holdout_s711–s716. The heuris-

tic audits use hcrit_can_cold_s607–s616 and hcrit_can_full_s701–s710. All evaluations use eval.success_contract=strict; this avoids the legacy failure mode where non-empty environment info dictionaries could be misread as successful rollouts. Table 10 summarizes the source run directories for each result block.

9. Conclusion

TRAK-Traj adapts gradient-based attribution to the trajectory level for offline robot imitation data curation. In a 10-seed cold-start curated Can MH study, it significantly outperforms matched random pruning while preserving a simple workflow: score trajectories, prune, and retrain. The held-out mixed-quality audit shows why this should remain a scoped claim: trajectory length is stronger there, and TRAK does not beat random on mean success. The result is best read as a focused, reproducible audit showing that practical data curation needs paired testing, cheap-baseline audits, prune-ratio sensitivity, and broader validation across tasks and policy classes.

A. Statistical Procedure Details

All confirmatory tests operate on paired seed-level deltas $D_s = Y_s^{\mathrm{TRAK}} - Y_s^{\mathrm{Random}}$ for the 10 curated Can seeds. The paired t-test tests whether the mean delta is positive. The Wilcoxon signed-rank test uses the signed ranks of nonzero seed deltas. The sign test uses only the number of TRAK wins. The permutation test randomly flips the sign of each paired delta under the null that method labels are exchangeable within seed; we use 50K random sign-flip draws. The bootstrap confidence interval resamples the 10 seed-level deltas with replacement and reports the percentile interval over the mean. The Bayesian value in Table 2 is descriptive rather than the basis for thresholding; it summarizes posterior mass on a positive mean effect under a simple seed-level normal model.

B. Source Artifact Summary

The heuristic audit rows are not used as primary evidence because they answer a different question from the pre-specified random-pruning comparison. They are reported to make clear that loss, trajectory length, and TracIn-style scores are serious comparators for practical data curation.

References

Bae, J., Ng, N., Lo, A., Ghassemi, M., and Grosse, R. B. If influence functions are the answer, then what is the question? In Advances in Neural Information Processing Systems, 2022.

Basu, S., Pope, P., and Feizi, S. Influence functions in deep learning are fragile. In International Conference on Learning Representations, 2021.

Belkhale, S., Cui, Y., and Sadigh, D. Data quality in imitation learning. In Advances in Neural Information Processing Systems, 2023.

Benavoli, A., Corani, G., Demšar, J., and Zaffalon, M. Time for a change: A tutorial for comparing multiple classifiers through bayesian analysis. Journal of Machine Learning Research, 18(77):1–36, 2017.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1):289–300, 1995.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. In Robotics: Science and Systems, 2023.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, 2 edition, 1988.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4RL: Datasets for deep data-driven reinforcement learning. arXiv preprint arXiv:2004.07219, 2020.

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning. In International Conference on Machine Learning, 2019.

Hejna, J., Bhateja, C. A., Jiang, Y., Pertsch, K., and Sadigh, D. ReMix: Optimizing data mixtures for large scale imitation learning. In Conference on Robot Learning, 2025.

Hoque, R., Mandlekar, A., Garrett, C., Goldberg, K., and Fox, D. IntervenGen: Interventional data generation for robust and data-efficient robot imitation learning. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2024.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., and Madry, A. Datamodels: Understanding predictions with data and data with predictions. In International Conference on Machine Learning, 2022.

Table 10. Primary artifact sources. Each run directory contains a metrics.jsonl file with baseline and pruned evaluation records.

Result	Seeds	Run-name pattern
Curated Can, primary	607–616	can_cold_can_mh_s*
Mixed-quality scan	701–710	can_full_can_mh_s*
Mixed-quality holdout	711–716	can_full_holdout_s*
Curated heuristic audit	607–616	hcrit_can_cold_s*
Mixed heuristic audit	701–710	hcrit_can_full_s*

Killamsetty, K., Durga, S., Ramakrishnan, G., and De, A. Grad-match: Gradient matching based data subset selection for efficient deep model training. In International Conference on Machine Learning, 2021.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In International Conference on Machine Learning, 2017.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In International Conference on Learning Representations, 2022.

Kress-Gazit, H., Hashimoto, K., Kuppuswamy, N., Shah, P., Horgan, P., Richardson, G., Feng, S., and Burchfiel, B. Robot learning as an empirical science: Best practices for policy evaluation. arXiv preprint arXiv:2409.09491, 2024.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In Advances in Neural Information Processing Systems, 2020.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643, 2020.

Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., and Martín-Martín, R. What matters in learning from offline human demonstrations for robot manipulation. In Conference on Robot Learning, 2021.

Mandlekar, A., Nasiriany, S., Wen, B., Akinola, I., Narang, Y., Fan, L., Zhu, Y., and Fox, D. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In Conference on Robot Learning, 2023.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. Foundations and Trends in Robotics, 7(1-2):1–179, 2018.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. Trak: Attributing model behavior at scale. In International Conference on Machine Learning, 2023.

Pomerleau, D. A. ALVINN: An autonomous land vehicle in a neural network. In Advances in Neural Information Processing Systems, 1989.

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. Estimating training data influence by tracing gradient descent. In Advances in Neural Information Processing Systems, 2020.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In International Conference on Artificial Intelligence and Statistics, 2011.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In International Conference on Learning Representations, 2018.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Conference on Empirical Methods in Natural Language Processing, 2020.

Toneva, M., Sordoni, A., Combes, R. T. d., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. In International Conference on Learning Representations, 2019.

Yoon, J., Arik, S. O., and Pfister, T. Data valuation using reinforcement learning. In International Conference on Machine Learning, 2020.

Zhao, T. Z., Kumar, V., Levine, S., and Finn, C. Learning fine-grained bimanual manipulation with low-cost hardware. In Robotics: Science and Systems, 2023.

Zheng, X., Pang, T., Du, C., Jiang, J., and Lin, M. Intriguing properties of data attribution on diffusion models. In International Conference on Learning Representations, 2024.