TASK CONDITIONS TELL MORE THAN THE TASK DATASET-PROPERTY LEAKAGE AND PRIVATE RELEASE IN WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robot policies often compress private demonstrations into reusable task conditions. Even when trajectories remain hidden, a released condition may expose a dataset property such as collection-domain composition. We introduce TASKLEAK, a release-time framework that measures dataset-property leakage against adaptive attackers and tests whether protected conditions still control robot behavior. Across multiple manipulation tasks, attacks reliably distinguish and estimate demonstration mixtures, and the signal persists across visual representations and an implicit VLA state. Common perturbation and linear erasure defenses either remain vulnerable after attacker retraining or damage control utility. We therefore construct a record-level differentially private release constrained by public task semantics. It reduces adaptive inference to near-chance behavior while preserving closed-loop task performance. Matched-state interventions further show that the protected condition retains task-specific control rather than merely accompanying success from the policy prior. Together, these results identify a privacy boundary in robot task conditioning and show that formal protection can coexist with behavioral utility.

1 INTRODUCTION

A robot can use a private demonstration set without receiving the demonstrations themselves. Some systems compress the set into a task condition that is stored, shared, or passed to a policy at execution time (James et al., 2018; Shen et al., 2026; Yang et al., 2026). The data owner can then share the ability to perform a task while keeping the source trajectories inaccessible. Yet the condition is computed from those trajectories. The same manipulation can be demonstrated in nominal and randomized domains. An observer of the released condition may infer the dataset's nominal-domain fraction property even though no source episode is visible (Fig. 1).

Although the two domains are not inherently personal, their mixture matters when a data owner intends to keep its collection strategy private. For example, how heavily a task was prepared with randomized demonstrations. Our simulator domains provide a controlled instance of that setting, rather than evidence of harm to an individual. The attacker knows how conditions are constructed and has labeled examples from the same task but cannot inspect the target collection. Under this access model, the fraction can disclose a choice about data preparation that was not intended for public sharing.

Property inference from aggregate data is already established (Ateniese et al., 2015; Suri & Evans, 2022; Kerkouche et al., 2023). The question here is what happens when the aggregate is also the signal that directs a robot. It crosses a data-release boundary and must remain useful to the policy. Hiding the dataset property by making the condition meaningless would leave that problem unsolved. Our attack estimates a collection proportion rather than the presence of one episode (Shokri et al., 2017; Pyrgelis et al., 2018). The Method section derives when averaging can strengthen this signal under explicit assumptions. That calculation alone cannot tell whether the signal survives a learned visual encoder, limited auxiliary data, or an adaptive defense. It also says nothing about whether a protected condition still guides action. Those questions require evidence at the interface the robot actually uses.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

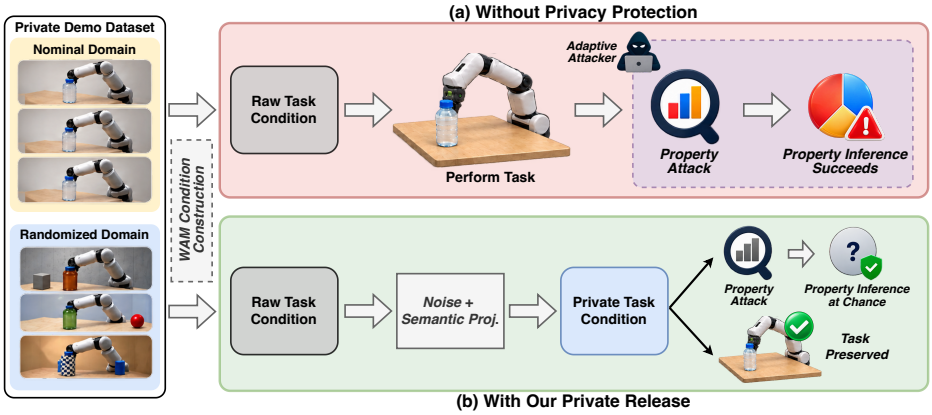


Figure 1: Task conditions can expose dataset properties. **(a) Without protection.** An adaptive attacker infers the private dataset's domain composition from the raw condition without viewing its demonstrations. **(b) With TASKLEAK.** The protected condition suppresses the tested property inference while preserving task-directed robot execution.

Two misleading outcomes follow from this dual role. Representation sanitization and concept erasure offer candidate defenses (Coavoux et al., 2018; Ravfogel et al., 2020; Belrose et al., 2023). A defense can seem private because it changes the representation seen by a fixed attacker, even though an attacker retrained on protected conditions recovers the signal. A robot can seem to retain utility because it succeeds from the scene or its policy prior, even if it ignores the protected condition. We address the first case with adaptive attacks and the second by changing only the condition from a matched initial state. A useful release must withstand the tested attacks while still making the robot respond to the intended task.

We study this question at the point where a task condition leaves the private demonstration collection. Our main setting is a demonstration-conditioned world action model whose task vector can actually be released and inspected (Yang et al., 2026). It gives us a concrete attacker observable and a policy that consumes the same vector. We also examine other representations and an internal state of a language-conditioned robot policy (Brohan et al., 2023b; Kim et al., 2024). These extensions test whether the dataset-composition signal is confined to one visual summary. These interfaces have different observables and do not share a single privacy unit or formal guarantee.

We organize TASKLEAK around the release path in Fig. 2. We use episode-disjoint examples to test whether a released condition carries the dataset-property signal. Attackers then retrain after each defense (Nasr et al., 2019). Finally, we construct a record-level differentially private condition and test it in paired robot rollouts. The guarantee bounds the effect of replacing one demonstration (Dwork et al., 2006; Dwork & Roth, 2014). Property inference and task execution are measured separately, since neither establishes the other.

Our experiments find this dataset-property leakage across tasks and representation interfaces. Larger aggregates improve discrimination in our setting. Common defenses leave leakage or impair control after attacker retraining. A release constrained by public task semantics reduces tested leakage while retaining closed-loop behavior. Wrong-task interventions show that this behavior remains condition dependent. Our formal privacy claim applies to the evaluated explicit release.

The contributions are threefold.

- We establish dataset-composition leakage from a task condition consumed by a robot policy. Analysis explains when aggregation sharpens the signal, while controls probe episode overlap, attack-specific artifacts, and dependence on one representation.
- We evaluate defenses against attackers retrained on the defended condition and construct a record-level private release that also preserves closed-loop control. This joins a formal guarantee with an empirical test of the dataset property and the robot's behavior.

- We use paired condition changes to show that successful control still depends on task information. Additional representation and policy-state studies locate where similar dataset-property signals remain observable beyond the explicit task vector.

2 Background and Related Work

Demonstration-conditioned control and WAMs. Task embeddings have long enabled few-shot control from visual demonstrations (James et al., 2018). BC-Z conditions broad imitation policies on language or human videos (Jang et al., 2022), while WAMs couple an imagined visual future to action generation (Shen et al., 2026). Generalist policies such as RT-1, RT-2, OpenVLA, and Octo condition control on language, images, or learned tokens (Brohan et al., 2023a;b; Kim et al., 2024; Octo Model Team et al., 2024), but standard OpenVLA does not consume a private demonstration set. ICRT, Vid2Robot, and Zero-WAM instead provide native demonstration-conditioned transformer interfaces (Fu et al., 2024; Jain et al., 2024; Zhou et al., 2026). Reusable visual and action representations offer additional comparison points (Nair et al., 2022; Majumdar et al., 2023; Chi et al., 2023; Zhang et al., 2026; Dexmal, 2026). We use LiLa-WAM (Yang et al., 2026) as the controlled case because its explicit visual-transition condition makes the release boundary observable, then use native or official representations only to test how far the phenomenon extends. Our target is the information exposed by task conditioning, rather than imagined-future integrity or action correctness (Peng et al., 2026a; Ma et al., 2026).

Property inference and aggregate leakage. Property inference recovers population-level facts rather than the presence of one record. Early meta-classifiers exposed training-set statistics (Ateniese et al., 2015). Later work attacked neural parameters (Ganju et al., 2018), collaborative updates (Melis et al., 2019), and black-box multi-party models (Zhang et al., 2021). Suri and Evans formalize this family as distribution inference and explicitly study mixture ratios (Suri & Evans, 2022). Aggregate releases can also leak membership (Pyrgelis et al., 2018) and dataset properties (Kerkouche et al., 2023). Recent VLA studies perform membership inference from policy behavior or internal signals (Peng et al., 2026b; Luan et al., 2026). This threat is distinct from model inversion (Fredrikson et al., 2015) and classic membership inference from black-box outputs (Shokri et al., 2017). Subsequent work connects membership risk to overfitting (Yeom et al., 2018), emphasizes low-FPR evaluation (Carlini et al., 2022), and strengthens white-box attacks (Nasr et al., 2019). We target neither reconstruction nor the membership of one trajectory, but these results motivate an adaptive, low-FPR privacy evaluation. Broader work frames privacy–utility coupling across the embodied-AI lifecycle and surveys world-model security threats (Fan et al., 2026; Liu et al., 2026), but does not evaluate a released multi-demonstration task summary. Our observable is a released multi-demonstration task condition rather than a trained weight, gradient, or policy output. The statistical possibility of aggregate property inference is established. The open question is whether this particular release exposes a dataset property, such as domain composition, and whether it can be protected while remaining useful for robot control.

Representation privacy and concept erasure. Per-example embeddings can leak input attributes, membership, and reconstructive information (Song & Raghunathan, 2020). Learned private representations expose a privacy–utility trade-off under adaptive attribute predictors (Coavoux et al., 2018). Our object is instead an aggregate of embeddings. INLP iteratively removes the row spaces of linear concept probes (Ravfogel et al., 2020). LEACE gives a closed-form minimum-distortion linear eraser (Belrose et al., 2023). Recent embedding defenses use concept-aware anisotropic mechanisms (Tsai et al., 2026), while subspace analyses warn that fitted erasure may not generalize (Naowarat et al., 2026). We claim none of these mechanisms as new. We study how they behave when the released object is a WAM task condition and the attacker is retrained after sanitization.

Differentially private aggregates. Clipping and Gaussian perturbation give record-level differential privacy for bounded vector means, and data-independent post-processing preserves the guarantee (Dwork et al., 2006; Dwork & Roth, 2014; Balle & Wang, 2018). DP-SGD demonstrates privacy-preserving deep learning (Abadi et al., 2016), and Rényi DP supports tight composition accounting (Mironov, 2017). Our contribution is not a new DP theorem. We adapt the mechanism to WAM conditions and show that semantic post-processing can recover closed-loop utility that direct Gaussian perturbation loses.

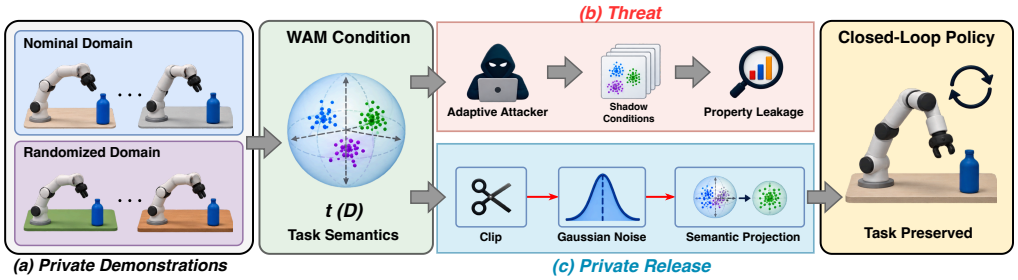


Figure 2: The task-condition privacy boundary. **(a)** Nominal- and randomized-domain demonstrations form a WAM condition. **(b)** An adaptive attacker infers the dataset's domain composition using shadow conditions. **(c)** Clipping, Gaussian noise, and public semantic projection protect the released condition before closed-loop control.

3 Problem Formulation

Let $H_\theta(D, x)$ denote the condition-dependent state used by a robot policy on a demonstration collection D and query state x. Depending on the system, this state can be an exported vector, a prompt-dependent transformer state, or an action distribution evaluated on a fixed public probe (Yang et al., 2026; Fu et al., 2024; Jain et al., 2024; Kim et al., 2024). A dataset property $q(D)$ creates a privacy channel only when the corresponding displacement in an attacker-observable H_θ is nonzero, following the distribution-inference view of dataset-level secrets (Suri & Evans, 2022; Zhang et al., 2021). Our experiments instantiate the strongest release-time case, where the state itself is explicitly exported.

For episode i, let s_i and g_i be its start and goal observations and ϕ a frozen visual encoder. LiLa-WAM (Yang et al., 2026) constructs transition features

$$z_i = \phi(g_i) - \phi(s_i), \qquad t(D) = \frac{1}{n}\sum_{i=1}^{n} z_i. \tag{1}$$

The task condition $t(D) \in \mathbb{R}^{1024}$ is the released observable. Each episode belongs to a nominal domain ($A = 1$) or a randomized domain ($A = 0$). Our main dataset property is the nominal fraction p in D. These domains are a controlled proxy for a collection strategy. The fraction is confidential only when the data owner does not intend to disclose that strategy.

Adversary. The adversary observes $t(D)$ and has a small auxiliary pool of labeled nominal-domain and randomized-domain episodes from the same task family. It knows the condition recipe and any defense. It does not see the target episodes. For defended releases, it can create and sanitize shadow conditions before fitting its attack. This mechanism-aware access follows adaptive inference evaluations that retrain the adversary on the released representation (Coavoux et al., 2018; Nasr et al., 2019). The main threat model does not assume access to arbitrary internal activations. Such access defines a distinct implicit-condition interface and must be tested on a model that natively consumes demonstrations.

4 Method

4.1 Mixture-Property Inference

Let domain means be μ_0, μ_1, let $\delta = \mu_1 - \mu_0$, and let domain covariances be Σ_0, Σ_1. This mixture view follows distribution and aggregate property inference (Suri & Evans, 2022; Kerkouche et al., 2023). For a fixed-count aggregate of independent episodes,

$$\mathbb{E}[t \mid p] = \mu_0 + p\delta, \tag{2}$$

$$\mathrm{Cov}(t \mid p) = \frac{(1-p)\Sigma_0 + p\Sigma_1}{n}. \tag{3}$$

Thus the property creates a linear displacement while aggregation suppresses episode noise as $n^{-1/2}$. For direction w, we estimate

$$\hat{p}(t) = \frac{w^\top (t - \hat{\mu}_0)}{w^\top (\hat{\mu}_1 - \hat{\mu}_0)}. \tag{4}$$

Our endpoint attack uses a diagonal-whitened approximation to $w \propto \Sigma^+ \delta$. Under a Gaussian approximation, standardized separation grows as $\sqrt{n}$.

Aggregation law and implicit conditions. Under the shared-covariance model $t \mid p \sim \mathcal{N}(\mu_0 + p\delta, \Sigma/n)$, the optimal linear score for fractions p_0 and p_1 has

$$d_n = |p_1 - p_0| \sqrt{n\, \delta^\top \Sigma^+ \delta}, \qquad \mathrm{AUC} = \Phi(d_n/\sqrt{2}). \tag{5}$$

Thus aggregation can make this dataset property easier, rather than harder, to distinguish. This statement is architectural only through the observable. If an implicit policy state is locally $H_\theta(D, x) = F_x(t(D))$ with Jacobian J_x, its first-order property direction is $J_x \delta$ and its covariance is $J_x \Sigma J_x^\top / n$. The same scaling survives only when this direction is not annihilated and the resulting state or output is available to the attacker. This gives a testable bridge to transformer conditioners (Fu et al., 2024; Jain et al., 2024; Kim et al., 2024). It does not assert that language-only VLAs expose LiLa's multi-demonstration privacy unit.

We additionally train three shadow-aggregate classifiers comprising L2 logistic regression, a 300-tree random forest, and a one-hidden-layer MLP. Standardization and PCA-64 are fit only on attacker-role aggregates. Predicted class probabilities over the four mixture fractions are converted to a scalar by taking their expected fraction. All tasks share the same hyperparameters. This multi-attack design follows the principle that property leakage should not be identified through one privileged probe (Ganju et al., 2018; Zhang et al., 2021).

4.2 Defense-aware evaluation

Let M denote any sanitizer. Evaluating a classifier trained on raw conditions against $M(t)$ confounds representation shift with privacy. We instead give the attacker the mechanism and form sanitized shadows $\{M(t(D_j^{\mathrm{aux}}))\}_j$. All preprocessing, dimensionality reduction, and prediction are refit on these shadows before evaluation on disjoint target aggregates. For stochastic mechanisms, defender, attacker, and evaluation roles also use independent randomness. We regard a mechanism as private against the tested family only when every attack and split remains below the stated AUC threshold. Utility is assessed separately. This minimax-style protocol does not assume our attack family is optimal, but avoids crediting a defense for breaking one stale probe, a concern shared by adaptive representation and privacy evaluations (Coavoux et al., 2018; Carlini et al., 2022).

4.3 Private task-condition release

The release mechanism operates on transition vectors rather than on their unbounded mean. From a public/reference anchor a, estimated once from a reference split disjoint from protected target members, clip each transition as

$$c_i = a + (z_i - a) \min\left(1, \frac{C}{\|z_i - a\|_2}\right), \qquad \bar{c} = \frac{1}{n} \sum_i c_i. \tag{6}$$

Replacement of one trajectory changes $\bar{c}$ by at most $2C/n$. We release

$$\widetilde{t} = \Pi_{\mathcal{B}(a, C/\sqrt{n})} \left(\bar{c} + \eta\right), \qquad \eta \sim \mathcal{N}(0, \sigma_{\epsilon,\delta}^2 I), \tag{7}$$

where $\sigma_{\epsilon,\delta}$ is calibrated to this sensitivity and $\delta = 10^{-5}$ using the Gaussian mechanism (Dwork et al., 2006; Balle & Wang, 2018). The ball center and radius are fixed without target members. Deployment therefore requires an actually non-private reference source rather than relabeling protected examples as public. Projection is deterministic post-processing, so the entire release remains (ϵ, δ)-DP (Dwork & Roth, 2014). Geometrically, it prevents high-dimensional noise from placing a condition far outside the task scale.

The mechanism has two deliberately separate design constraints. Clipping and Gaussian calibration provide the record-replacement guarantee, whereas the public semantic ball constrains control

displacement without consuming privacy. We select the mapping using defender data and four defense-aware nonlinear attack families, then hold it unchanged for closed-loop evaluation. Repeated or overlapping releases require composition. Neither the public projection nor empirical attack failure changes that accounting (Mironov, 2017).

Guarantee. Consider neighboring collections $D \simeq D'$ of equal size that differ in one trajectory, with a, C, and the projection radius fixed independently of both collections. Because every clipped vector lies in $\mathcal{B}(a, C)$, $\|\bar{c}(D) - \bar{c}(D')\|_2 \leq 2C/n$. Calibrating $\sigma_{\epsilon,\delta}$ to this sensitivity makes the noisy mean (ϵ, δ)-DP under replacement adjacency. Applying $\Pi_{\mathcal{B}}$ and the downstream WAM are data-independent mappings of that release, so both inherit the same guarantee by post-processing. We additionally compute the exact equal-covariance Gaussian hockey-stick divergence for all observed source rows. This certificate checks the implementation but is not needed for the proof.

Why the semantic ball helps. Raw isotropic noise spends magnitude in all 1,024 coordinates, although valid task conditions occupy a much smaller region around the public task prototype. The projection removes only the radial excess of the already private vector. For $v = \bar{c} + \eta - a$,

$$\Pi_{\mathcal{B}(a,r)}(a + v) = a + v \min(1, r/\|v\|_2), \qquad r = C/\sqrt{n}. \tag{8}$$

It uses neither the target collection nor a learned private residual. Thus its role is not to add privacy, but to encode a public validity constraint that restores a plausible scale for control. The radius is chosen before evaluation. Selecting it on target success would invalidate the protocol even though DP itself would still hold for a separately public choice.

Behavioral estimand. Privacy is useful only if the released condition causally affects control. For each task, condition seed, and simulator seed, we hold the initial scene fixed and change only the condition. The primary utility estimand is the paired success difference between the private release and aligned raw control. Wrong and zero conditions test whether the policy responds to task semantics and rule out an unconditional-policy explanation in LiLa-WAM (Yang et al., 2026). Because condition-seed replicates share simulator seeds, uncertainty resamples both crossed factors rather than treating episodes as independent.

5 Experiments

Benchmarks and tasks. Our primary benchmark combines LiLa-WAM with RoboTwin 2.0 (Yang et al., 2026; Chen et al., 2025). We evaluate `adjust_bottle`, `click_bell`, `click_alarmclock`, and `turn_switch`, holding out the last task during initial attack development. Each task contains nominal-domain and randomized-domain expert episodes, with exact source counts reported in Appendix A. Model-generalization experiments additionally use BridgeData V2 with OpenVLA (Walke et al., 2024; Kim et al., 2024) and official representations from OpenCLAP and DW05 (Zhang et al., 2026; Dexmal, 2026).

Experimental setup. We reproduce the published 1024-D DINOv2 task conditions (Oquab et al., 2023; Yang et al., 2026) and construct without-replacement aggregates of size $\{20, 40, 100\}$. Attack evaluation uses episode-disjoint auxiliary and target halves, while defense experiments reserve separate thirds for defense fitting, adaptive-attacker fitting, and evaluation. Every attacker is retrained after each transformation, and no episode identity crosses roles within a split. We repeat the partitioning with three seeds. The private-release study uses size-40 conditions, four privacy budgets, three mechanism replicates, and ten paired simulator seeds per task. Appendix A gives the complete sampling rules and decision criteria.

Metrics and uncertainty. The primary discrimination problem separates 0% from 10% nominal-domain mixtures, while calibration estimates fractions in $\{0, 5, 10, 20\}\%$. We report ROC AUC, TPR at an empirical 1% FPR, FPR at 95% TPR, and fraction MAE, following the emphasis on low-FPR privacy evaluation in prior inference work (Carlini et al., 2022). Closed-loop utility is measured by paired task success and early-action agreement with the aligned raw condition. Confidence intervals use 1,000 stratified bootstrap resamples, with crossed resampling over condition and simulator seeds when both vary.

Table 1: Dataset-property inference for nominal-domain composition across aggregate sizes, including a low-data setting with 2 nominal and 20 randomized auxiliary episodes. Values average three episode-disjoint splits.

Task	AUC@20	TPR@1%@20	AUC@100	Aux 2/20 AUC
Adjust bottle	0.999	0.987	1.000	1.000
Click bell	0.999	0.983	1.000	1.000
Click alarmclock	0.939	0.507	1.000	0.953
Turn switch	0.995	0.921	1.000	1.000

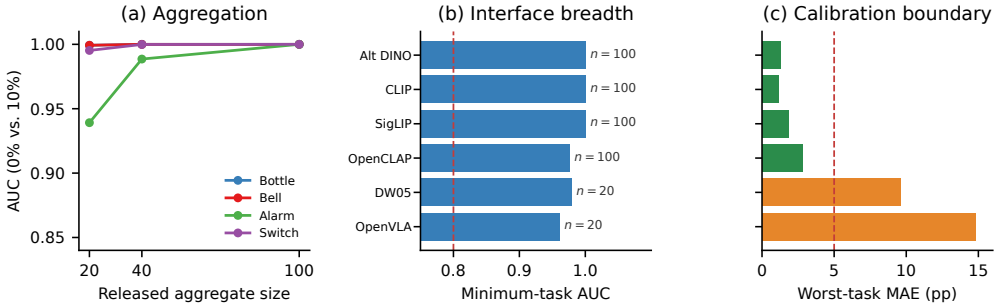


Figure 3: Dataset-property leakage under aggregation and across interfaces. **(a)** LiLa's 0%-versus-10% mixture AUC rises with aggregate size. **(b)** The weakest task remains distinguishable across six interfaces, with sample sizes beside the bars. **(c)** Mixture error meets the five-point target for alternate DINO, CLIP, SigLIP, and OpenCLAP, but not DW05 or OpenVLA. Sample sizes are protocol specific.

5.1 Dataset-Property Leakage

Leakage and aggregation. At aggregate size 100, the episode-disjoint attack reaches AUC 1.000 on all four tasks and reduces mixture error well below the constant-prior baseline. Leakage remains strong at size 20 and with only 2 nominal-domain plus 20 randomized-domain auxiliary episodes. Appendix Fig. 14 shows that aggregation especially strengthens low-FPR detection on alarmclock, whereas label-permutation controls remain at chance. These results establish a robust same-task threat rather than a task-agnostic attacker, consistent with the role of auxiliary distribution knowledge in property inference (Suri & Evans, 2022). Appendix Table 4 gives the task-level results and finite-population controls.

Attack robustness and transfer. Endpoint, logistic, forest, and MLP attacks give the same qualitative result (Fig. 9), so leakage is not an artifact of one estimator or attack family (Ganju et al., 2018; Zhang et al., 2021). A three-task attacker often transfers discrimination to a held-out task, but its fraction error rises to 5.11–11.19 points. Cross-task detection is therefore possible, whereas calibrated recovery still requires task-local auxiliary data. This distinction matters operationally because detecting a collection shift is easier than estimating its prevalence accurately.

5.2 Property and Model Generalization

Transfer summary. Alternate DINOv2, CLIP, SigLIP, and OpenCLAP support both discrimination and fraction recovery, whereas DW05 supports only the former (Oquab et al., 2023; Radford et al., 2021; Zhai et al., 2023; Zhang et al., 2026; Dexmal, 2026). Base OpenVLA's fused prefix reaches 0.977/0.961 mean/minimum-task AUC on 600 session-disjoint BridgeData V2 trajectories, although its fraction estimates remain poorly calibrated (Figs. 3 and 7) (Kim et al., 2024; Walke et al., 2024). The shared discrimination trend shows that the privacy channel is not unique to LiLa's native DINOv2 summary, while calibration remains interface dependent.

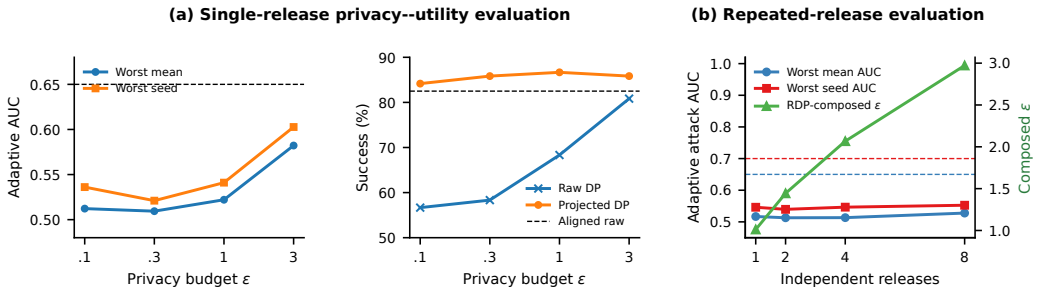


Figure 4: Privacy and utility of the semantic-ball release. **(a)** Projection keeps adaptive attack AUC below its threshold and restores success lost by raw Gaussian noise. **(b)** Attack AUC stays below the empirical thresholds through eight releases as the composed privacy budget rises.

5.3 Private Release and Behavioral Evaluation

Defense baselines. No tested baseline yields a joint privacy–utility operating point. Gaussian noise below $\sigma = 64$ remains detectable after attacker retraining, while $\sigma = 64$ suppresses the attack only by materially changing policy actions. LEACE and INLP preserve high feature similarity but leave a strong adaptive signal (Table 6) (Ravfogel et al., 2020; Belrose et al., 2023). Raw Gaussian-DP vectors suppress the tested attacks yet lose closed-loop success, showing that representation similarity and a DP guarantee do not by themselves imply behavioral utility. This motivates treating privacy and control as joint requirements rather than selecting a defense from representation metrics alone.

Certified private release. Semantic-ball projection resolves this failure (Table 2). Across four retrained attacks, its worst task/seed AUC remains below 0.65/0.70, while all four privacy budgets match or exceed the aligned raw success rate. The same pattern holds across 480 projected-DP episodes and a sequential-policy comparison. An independent Gaussian calculation verifies the target privacy level under the calibrated mechanism (Dwork et al., 2006; Balle & Wang, 2018). The guarantee applies to one clipped size-40 release, while a public anchor is reported separately and repeated releases require composition.

Repeated releases. Retrained RF, MLP, ExtraTrees, and histogram-gradient-boosting attacks remain near chance through eight independent ϵ=1 releases (Fig. 4). This favorable empirical result does not preserve the single-release guarantee. Gaussian-RDP composition reaches $\epsilon = 2.97$ at $K = 8$ (Mironov, 2017). Repeated exposure therefore still requires accounting and rate limiting even when the tested attacks remain weak.

Condition semantics. With observations and flow noise shared, private first-action chunks stay within 0.41% RMS of raw. Wrong-task and zero controls change action identity. Both clustered margins are positive on all four tasks (Appendix Fig. 12). These matched-state comparisons isolate the condition intervention before trajectories diverge and support genuine task-semantic use by LiLa-WAM (Yang et al., 2026).

Table 2: Privacy and closed-loop control after semantic-ball release over 120 episodes. Δ is measured from aligned raw success (82.5%).

ϵ	worst mean AUC	worst seed AUC	cosine	raw DP success	projected DP success	Δ raw (pp)
0.1	.512	.536	.962	56.7	84.2	+1.7
0.3	.509	.521	.962	58.3	85.8	+3.3
1.0	.522	.541	.963	68.3	86.7	+4.2
3.0	.582	.603	.964	80.8	85.8	+3.3

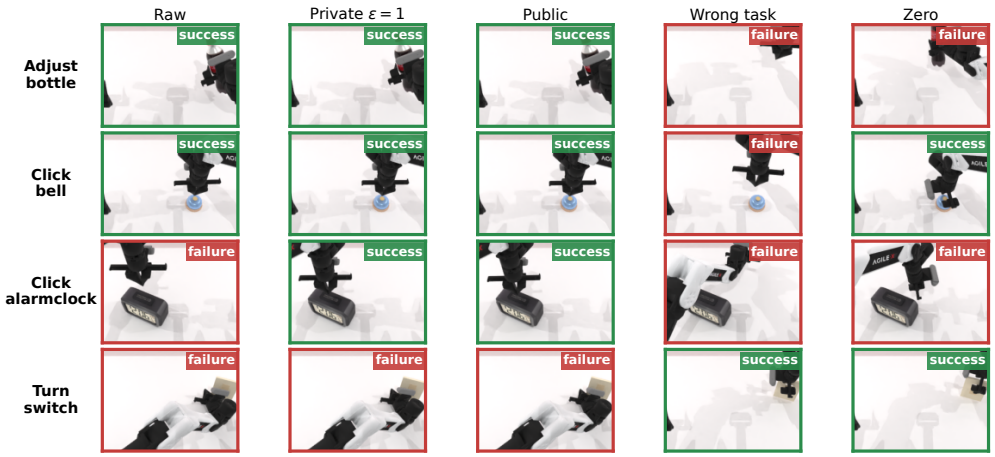


Figure 5: LiLa-WAM rollouts under raw, private, public, wrong-task, and zero conditions. Each row shares an initial state and simulator seed. Borders mark the recorded outcomes. Quantitative comparisons include every rollout.

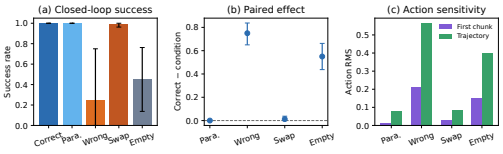


Figure 6. OpenVLA-OFT responds to instruction semantics. Correct and paraphrased prompts preserve success, whereas wrong and empty prompts alter both the first action and the ensuing trajectory.

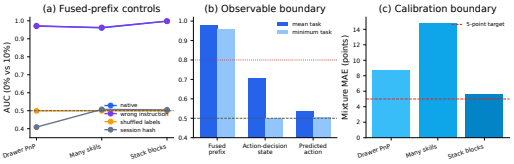


Figure 7. Collection environment is visible in OpenVLA's fused prefix on BridgeData V2. The signal persists after instruction permutation but vanishes in label, session, and final-action controls.

5.4 Causal Condition Use Across Architectures

Transformer-VLA intervention. Across 400 matched-state OpenVLA-OFT rollouts, correct/paraphrased prompts reach 100% success versus 25/45% for wrong/empty prompts, which also alter first actions and trajectories (Fig. 6). This verifies native instruction use in OpenVLA-OFT (Kim et al., 2024). The intervention rules out an explanation based only on shared initial observations, without extending LiLa's DP guarantee.

Evidence from real-robot data. Across 600 session-disjoint BridgeData V2 trajectories, base OpenVLA's fused prefix separates collection environments with mean/minimum-task AUC 0.977/0.961. Label and session controls remain near chance, and action outputs lack the signal, localizing leakage to an implicit VLA state (Kim et al., 2024; Walke et al., 2024). The separation between fused prefix and action outputs narrows the observable boundary, but this offline result does not establish real-robot control or calibrated mixture recovery.

6 Conclusion

Robot task conditions can reveal dataset properties while the source demonstrations remain hidden. We introduced TaskLeak to study this release boundary through dataset-property inference, adaptive defense tests, and behavioral interventions. The composition signal appears across tasks and representation interfaces, while standard perturbation and erasure fail to preserve both privacy and control. Our semantically constrained private release suppresses tested attacks and preserves closed-loop behavior. Paired LiLa and OpenVLA interventions show that task information remains behaviorally relevant. Broader representation and VLA evidence suggests that the concern extends beyond the explicit encoder, while the formal guarantee applies to the evaluated LiLa release. These results provide a path to privacy-aware task conditioning with verified control utility.

AI USE STATEMENT

Generative AI tools were used for literature and experiment organization, assistance with analysis and plotting code, debugging, LaTeX restructuring, and language editing. They were not used as a source of empirical evidence or to fabricate, replace, or alter experimental observations. The authors inspected the cited literature, executed and tested the code, checked the reported numbers against the stored result artifacts, and reviewed all AI-assisted text and figures.

ETHICS STATEMENT

This work studies privacy and security risks in released robot-learning representations. It uses existing robotic demonstration data and simulator rollouts and does not involve human subjects or attempt to identify or reconstruct individuals. Nevertheless, property-inference methods can expose confidential information about data-collection pipelines, and repeated access can increase cumulative privacy loss. We report the attack to enable measurement and mitigation, evaluate adaptive defenses, avoid releasing private source episodes, and recommend data minimization, access control, rate limiting, and privacy accounting. Any deployment should respect dataset licenses, organizational policies, applicable law, and the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Section 5 specifies the datasets, aggregate sizes, attacker pools, privacy budgets, decision criteria, seeds, and closed-loop protocols. The appendix gives the complete split and sampling procedure, task-level results, defense grids, composition accounting, and additional interface evaluations. The supplementary source bundle includes the experiment settings.

REFERENCES

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM Conference on Computer and Communications Security*, pp. 308–318, 2016.

Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems*, 2023a.

Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *Conference on Robot Learning*, 2023b.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*, pp. 1897–1914, 2022.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *Robotics: Science and Systems*, 2023.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. Privacy-preserving neural representations of text. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1–10, 2018.

Dexmal. Dw05: A multimodal world model for embodied intelligence. OpenDW software release, https://github.com/dexmal/opendw, 2026.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284, 2006.

Xiaoliang Fan, Jiarui Chen, Zhuodong Liu, Ziqi Yang, Peixuan Xu, Ruimin Shen, Junhui Liu, Jianzhong Qi, and Cheng Wang. Position: Embodied ai requires a privacy-utility trade-off. *arXiv preprint arXiv:2605.05017*, 2026.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM Conference on Computer and Communications Security*, pp. 1322–1333, 2015.

Letian Fu, Huang Huang, Gaurav Datta, Lawrence Yunliang Chen, William Chung-Ho Panitch, Fangchen Liu, Hui Li, and Ken Goldberg. In-context imitation learning via next-token prediction. *arXiv preprint arXiv:2408.15980*, 2024.

Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *ACM Conference on Computer and Communications Security*, pp. 619–633, 2018.

Vidhi Jain, Maria Attarian, Nikhil J. Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R. Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *arXiv preprint arXiv:2403.12943*, 2024.

Stephen James, Michael Bloesch, and Andrew J. Davison. Task-embedded control networks for few-shot imitation learning. In *Conference on Robot Learning*, pp. 783–795, 2018.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pp. 991–1002, 2022.

Raouf Kerkouche, Gergely Ács, and Mario Fritz. Client-specific property inference against secure aggregation in federated learning. *Proceedings of the Workshop on Privacy in the Electronic Society*, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. OpenVLA: An open-source vision-language-action model. *Conference on Robot Learning*, 2024.

Fazhong Liu, Zhuoyan Chen, Haozhen Tan, Yan Meng, Guoxing Chen, and Haojin Zhu. Security of world-model-based embodied ai: A lifecycle of threats, defenses, and evaluation. *arXiv preprint arXiv:2607.28226*, 2026.

Xukun Luan, Jinyan Liu, Xuesong Li, Yuanguo Bi, Renjun Wu, Zhongxiang Lei, and Di Wang. Vlaleaks: Membership inference attacks against vision-language-action models. *arXiv preprint arXiv:2606.15165*, 2026.

Xiangkai Ma, Yue Ma, Junjie Wang, Sheng Xu, Mingyang Li, Han Zhang, Yuzheng Zhuang, Wenzhong Li, and Zhihao Yuan. Decoupling intention from trajectory: A representational deduction framework for world action models. *arXiv preprint arXiv:2608.06994*, 2026.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 2023.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *IEEE Symposium on Security and Privacy*, pp. 691–706, 2019.

Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium*, pp. 263–275, 2017.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pp. 892–909, 2022.

Burin Naowarat, Hao Tang, and Sharon Goldwater. A framework for analyzing concept representations in neural models. *arXiv preprint arXiv:2605.01381*, 2026.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy*, pp. 739–753, 2019.

Octo Model Team, Dibya Ghosh, Homer Walke, et al. Octo: An open-source generalist robot policy. *Robotics: Science and Systems*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Quanquan Peng, Yutong Liang, Rui Yan, Nicklas Hansen, and Xiaolong Wang. Fact: Failure-aware causal training for world-action models. *arXiv preprint arXiv:2608.10232*, 2026a.

Yuefeng Peng, Mingzhe Li, Kejing Xia, Renhao Zhang, and Amir Houmansadr. Membership inference attacks on vision-language-action models. *arXiv preprint arXiv:2605.07088*, 2026b.

Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. In *Network and Distributed System Security Symposium*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*, 2020.

Qiuhong Shen, Shihua Zhang, Yue Liao, Qi Li, Zhenxiong Tan, Shizun Wang, Shuicheng Yan, and Xinchao Wang. World action models: A survey. *arXiv preprint arXiv:2606.20781*, 2026.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *ACM Conference on Computer and Communications Security*, pp. 377–390, 2020.

Anshuman Suri and David Evans. Formalizing and estimating distribution inference risks. *Proceedings on Privacy Enhancing Technologies*, 2022(4):528–551, 2022.

Yu-Che Tsai, Hsiang Hsiao, Kuan-Yu Chen, and Shou-De Lin. Concept-aware privacy mechanisms for defending embedding inversion attacks. *arXiv preprint arXiv:2602.07090*, 2026.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale, 2024.

Fan Yang, Yuting Su, Xiaobo Wang, Yuncheng You, Fugui Fan, Yuting Wu, Minghui Wu, Chenxu Zhao, Jiahong Ning, and Peiguang Jing. Lila-wam: Lightweight latent reasoning world-action model for robotic manipulation. *arXiv preprint arXiv:2608.03701*, 2026.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, pp. 268–282, 2018.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Chubin Zhang, Jianan Wang, Zifeng Gao, Yue Su, Tianru Dai, Cai Zhou, Jiwen Lu, and Yansong Tang. Clap: Contrastive latent action pretraining for learning vision-language-action models from human videos. *arXiv preprint arXiv:2601.04061*, 2026.

Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. Leakage of dataset properties in multi-party machine learning. In *USENIX Security Symposium*, pp. 2687–2704, 2021.

Jiaming Zhou, Qihang Zhang, Gangwei Xu, Cunxin Fan, Yujie Zhao, Ruilin Wang, Yiming Luo, Shuai Yang, Xing Zhu, Yujun Shen, Junwei Liang, and Yinghao Xu. Zero-WAM: In-context world-action modeling from human videos for open-ended task generalization, 2026.

A Experimental Details

Splits and sampling. For every task and seed $\{20260915, 20260916, 20260917\}$, the two-way experiments independently permute nominal-domain and randomized-domain episodes and divide each domain into attacker-fit and evaluation halves. The defense experiments use three disjoint thirds for defense fitting, adaptive-attacker fitting, and evaluation. The original characterization queries sample with replacement only within their assigned role. The finite-population and object-property extensions sample without replacement inside each query. Each fraction in $\{0, 0.05, 0.10, 0.20\}$ has 500 evaluation queries. Learned attacks receive 250 shadow queries per fraction. No evaluation episode fits preprocessing, an attack, or a defense.

Uncertainty and operating points. Each reported confidence interval uses 1,000 stratified bootstrap resamples of the four fraction-specific query sets. TPR@1% FPR uses the higher 99th percentile of the zero-fraction scores. FPR@95% TPR uses the lower 5th percentile of the 10%-fraction scores. Since finite samples and strict inequalities are used, the observed FPR at the first threshold is reported alongside TPR in the machine-readable files.

Data cardinalities. The source counts for the nominal/randomized domains are 50/489 for adjust bottle, 50/500 for click bell, 50/500 for click alarmclock, and 48/484 for turn switch. The turn-switch departure from the initially expected 50/482 split was resolved before feature extraction by matching the official archive hash and total of 532 valid episodes. No record was added or removed after inspecting leakage.

B Task-Level Leakage Results

Table 3 reports the diagonal-whitened endpoint attack at aggregate size 100. The confidence interval column is the envelope of the three independently bootstrapped seed intervals. It is deliberately more conservative than presenting a pooled-query interval.

Table 3: Collection-composition inference across four RoboTwin tasks. MAE is reported in percentage points.

Task	Mean AUC	Min AUC	AUC CI envelope	MAE (pp)	TPR@1%	FPR@95%
Adjust bottle	1.0000	1.0000	[1.0000, 1.0000]	1.02	1.000	0.000
Click bell	1.0000	1.0000	[1.0000, 1.0000]	0.99	1.000	0.000
Click alarmclock	0.9989	0.9980	[0.9961, 1.0000]	2.85	0.965	0.006
Turn switch	1.0000	1.0000	[1.0000, 1.0000]	1.29	1.000	0.000

The nonlinear comparison uses StandardScaler and PCA-64 fit only on attacker shadows, followed by fixed multinomial logistic regression, a 300-tree random forest, or a width-64 one-hidden-layer MLP. Class probabilities are converted to expected mixture fractions. No validation split or task-wise hyperparameter selection is used.

C Finite-Population and Null-Control Evaluation

Sizes 20, 40, and 100 exactly represent every requested mixture fraction. Each query contains unique episodes, while independently generated queries may reuse episodes. Table 4 reports the endpoint estimator. The null control permutes binary query labels relative to the observed scores. Mean null AUC is 0.499–0.502. A separate episode-label diagnostic is preserved in the artifact but is not exchangeable because finite random group imbalance can be magnified by aggregation.

D Task-Held-Out Transfer and Published Vectors

For each target, PCA-64 logistic preprocessing and classification are fit on shadow conditions from the other three tasks only. Table 5 shows that discrimination often transfers, but calibrated recovery does not. All four mean MAEs exceed 5 percentage points. The final column applies the same

Table 4: Leakage persists when each condition contains unique episodes, whereas permuting query labels returns discrimination to chance.

Task	AUC@20	TPR@1%@20	AUC@100	MAE@100 (pp)	Null AUC
Adjust bottle	0.999	0.987	1.000	0.89	0.502
Click bell	0.999	0.983	1.000	0.87	0.500
Click alarmclock	0.939	0.507	1.000	2.63	0.500
Turn switch	0.995	0.921	1.000	1.13	0.499

Table 5: Leave-one-task-out discrimination, fraction recovery, and errors on the four published task vectors.

Held-out task	Mean AUC	Min AUC	TPR@1%	MAE (pp)	Pub. error (pp)
Adjust bottle	0.795	0.667	0.466	6.84	7.37
Click bell	1.000	1.000	1.000	5.11	10.29
Click alarmclock	1.000	1.000	1.000	5.80	7.31
Turn switch	0.997	0.994	0.935	11.19	10.98

task-disjoint models to the single official vector for each held-out task. It is an applicability diagnostic, not a population-level estimate of attack AUC.

We also evaluate source-configuration shift, six directed encoder/recipe shifts, pooled cross-task calibration, and attacker-shadow-only linear and isotonic recalibration. ExtraTrees and histogram gradient boosting are included alongside logistic, endpoint, RF, and MLP attacks. Every score, including weak transfer and poorly calibrated cells, is retained in the source- and representation-shift artifacts rather than selected after target evaluation.

E Additional Adaptive-Defense Results

All defenses are fit only on the defender role and then applied independently to attacker shadows and evaluation aggregates. Gaussian-4 uses diagonal population scale divided by $\sqrt{100}$. Raw and standardized projections use the same ranks on every task. LEACE uses pseudoinverse tolerance 10^{-8}. INLP runs exactly eight L2-logistic iterations. Table 7 aggregates all four tasks using worst-task AUC and average task-condition distortion. The privacy criterion requires both attacks to remain at mean AUC ≤ 0.65 and maximum-replicate AUC ≤ 0.70 on every task.

Every method exceeds at least one privacy threshold. We therefore do not spend closed-loop evaluation on directional, LEACE, or INLP mechanisms. None is a candidate defense whose utility remains unmeasured.

Additional defenses include defender-fit trimmed/winsorized aggregation and shrinkage, bottleneck adversarial sanitization, aggregate-level adversarial sanitization, conditional transport, and randomized sketch plus quantization. None removes the collection signal across all four tasks. The clipped Gaussian family with semantic-ball post-processing is the only method that supports both privacy and control, so it receives the full closed-loop and offline utility evaluation.

F Extended Adaptive Gaussian Frontier

The extended Gaussian study separates episodes from the nominal and randomized domains into disjoint defense-scale, attacker-fit, and evaluation thirds. Four attacks are retrained at each noise level. Table 8 reports the worst cell over four tasks and four attacks. Only $\sigma = 64$ suppresses all four attacks, but its action RMS is already unacceptable on both behavior tasks.

G Differentially Private Release and Closed-Loop Utility

For each task and mechanism seed, defender, attacker, and evaluation episodes are disjoint. The defender split fixes anchor a and median radius C. Each member is clipped into the radius-C ball

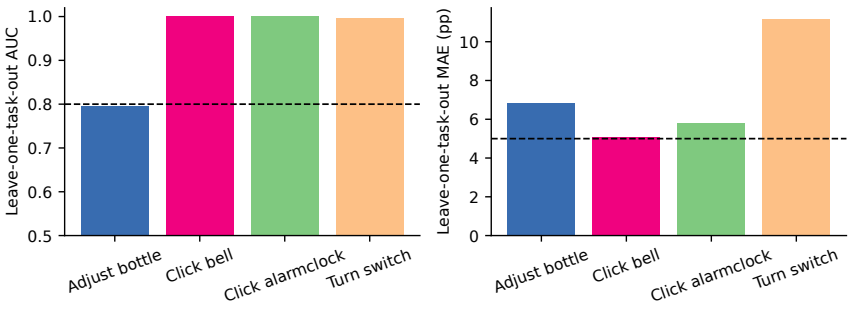


Figure 8: Task-held-out discrimination transfers unevenly, while fraction estimates exceed five percentage points of error on every target task.

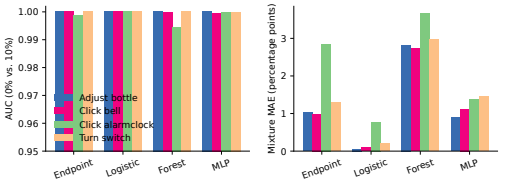


Figure 9. Leakage is consistent across endpoint, logistic, forest, and MLP estimators on all four tasks.

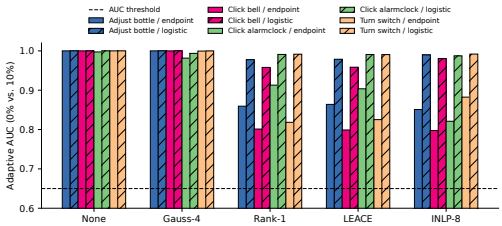


Figure 10. Linear erasure and perturbation leave substantial collection signal after the logistic attacker is retrained on transformed conditions (hatched bars).

about a. Hence replacing one member of a size-40 mean has sensitivity $2C/40$. Isotropic Gaussian noise uses $\sigma = (2C/40)\sqrt{2\log(1.25/\delta)}/\epsilon$ with $\delta = 10^{-5}$. An independent certificate additionally evaluates the exact equal-covariance Gaussian hockey-stick divergence. Its maximum over 96 source rows is 3.22×10^{-7}.

Directly perturbed vectors lose closed-loop utility at all four epsilon values that suppress the defense-aware attack suite. Before observing the remaining task outcomes, we specified four post-processors. The selected nonconstant mapping projects the already private vector into the radius-$C/\sqrt{40}$ ball about a. Attack AUC remains below the stated thresholds on every task under random forest, MLP, ExtraTrees, and histogram gradient boosting. Because the mapping and its parameters do not depend on target members, it preserves the input (ϵ, δ) guarantee.

All 60 task–budget–mechanism comparisons align to the raw controls on collection membership and ten expert-valid simulator seeds. Across 480 nonconstant private episodes, success remains within the prespecified pooled and per-task margins. A 120-episode constant public-anchor reference is reported separately and is not counted as a private aggregate. The guarantee covers one size-40 release under replacement adjacency. Repeated or adaptive releases require composition.

Aligned condition interventions. We reuse the raw, private-ϵ=1, and public-anchor evaluations above and add 24 closed-loop combinations comprising four tasks, three condition seeds, and wrong-task/zero conditions, each with the same ten expert-validated simulator seeds. Table 11 reports all 600 aligned episodes across the five conditions. Pooled success for raw, private, public-anchor, wrong-task, and zero conditions is 82.5%/86.7%/85.8%/10.8%/42.5%. The private condition exceeds the mean of wrong-task and zero controls by 60.0 points, with a positive difference on all four tasks. The contrast shows that success depends on the task condition rather than an unconditional policy prior.

To examine output-level preservation, we select 16 episodes from the randomized domain per task and four ordered time positions per episode, yielding 64 held-out policy queries per task. For three mechanism seeds and all four private epsilon levels, we compare the full 32×14 action chunks and all decoded future tokens against their aligned raw collection conditions using shared flow noise. The 48 task–budget–mechanism settings contain 3,072 private condition-query comparisons. The

Table 6. Task-level leakage and distortion for the strongest erasure baselines. Relative ℓ_2 is measured from the original condition.

Task	Defense	Endpoint AUC	Logistic AUC	Rel. ℓ_2
Adjust bottle	Rank-1	0.859	0.978	0.109
Adjust bottle	LEACE	0.864	0.979	0.070
Adjust bottle	INLP-8	0.851	0.990	0.545
Click bell	Rank-1	0.801	0.958	0.109
Click bell	LEACE	0.799	0.958	0.081
Click bell	INLP-8	0.797	0.980	0.225
Click alarmclock	Rank-1	0.913	0.991	0.096
Click alarmclock	LEACE	0.904	0.991	0.082
Click alarmclock	INLP-8	0.821	0.987	0.289
Turn switch	Rank-1	0.818	0.992	0.076
Turn switch	LEACE	0.825	0.990	0.049
Turn switch	INLP-8	0.882	0.992	0.291

Table 7. Privacy and representation distortion across all tested defenses. "Worst" denotes the maximum over four tasks.

Defense	Worst endpoint AUC	Worst logistic AUC	Worst seed logistic AUC	Mean rel. ℓ_2	Pass
None	1.000	1.000	1.000	0.000	no
Gaussian-4	1.000	1.000	1.000	0.511	no
Raw rank-1	0.913	0.992	0.999	0.098	no
Std. rank-1	0.913	0.992	0.999	0.098	no
Std. rank-4	0.958	0.995	0.999	0.114	no
Std. rank-8	0.948	0.996	0.999	0.121	no
LEACE	0.904	0.991	0.998	0.071	no
INLP-8	0.882	0.992	1.000	0.337	no

Table 8: Privacy–utility trade-off under Gaussian perturbation. Dashes mark noise levels for which behavior was not evaluated because attack AUC remained above the prespecified privacy threshold.

σ	Worst mean AUC	Worst seed AUC	Bell RMS (pp)	Alarm RMS (pp)	Joint pass
0	1.000	1.000	0.00	0.00	no
4	1.000	1.000	0.18	0.34	no
8	0.996	0.998	–	–	no
16	0.929	0.934	–	–	no
32	0.765	0.770	–	–	no
64	0.632	0.654	17.85	13.96	no

worst task mean action relative RMS is 0.414%, the worst seed mean is 0.466%, and the minimum query future cosine is 0.9973. Every task and privacy budget remains within the chosen action and future-token tolerances.

H Transformer and Early-Action Evidence

ICRT data suitability. We first screened the official ICRT-MT release because ICRT natively consumes sensorimotor demonstrations as transformer context. Its 1,098 trajectories cover 26 tasks, but the largest task contains only 84 trajectories. The study design requires 100 episode-disjoint demonstrations on each property side within at least three tasks. Even a hypothetical 50-per-side design would need 100 trajectories per eligible task, so the released data cannot support the test. Lowering the sample requirement after this census, or pooling tasks and confounding the property with task identity, would not provide a valid replication.

OpenVLA-OFT protocol. The language-conditioned replication uses the official OpenVLA-OFT LIBERO-Spatial checkpoint and task IDs 2, 6, 7, and 9. For each task, the 20 official initial states are serialized once and reused across correct, one meaning-preserving paraphrase, cyclic wrong-task, object-swapped, and empty instructions. Each condition receives a fresh environment instantiated from the same state. We retain every success outcome, first action chunk, normalized action/EEF trajectory, and rendering. No state is selected by outcome. Correct exceeds the pooled three negative conditions by 43.75 points, correct and paraphrase agree exactly, and the positive taskwise direction holds in 4/4 tasks. Wrong-task and empty success are 25% and 45%, whereas object swap is 98.75%. The latter is explicitly retained as a weak intervention rather than grouped into a stronger claim.

Offline BridgeData/OpenVLA protocol. We separately evaluate the official base OpenVLA checkpoint on BridgeData V2 real-robot trajectories (Kim et al., 2024; Walke et al., 2024). Before model inference, a complete 128-shard metadata census identifies environment family (toy kitchen versus non-kitchen) as the first property satisfying the eligibility criteria. The three eligible within-task counts are 241/106 for drawer PnP, 113/466 for many skills, and 141/166 for stack blocks. Every side has at least ten collection sessions and no session contributes more than 20%. A global session assignment then allocates 50 attacker-fit and 50 evaluation episodes per side and task, for 600 episodes with no session crossing roles.

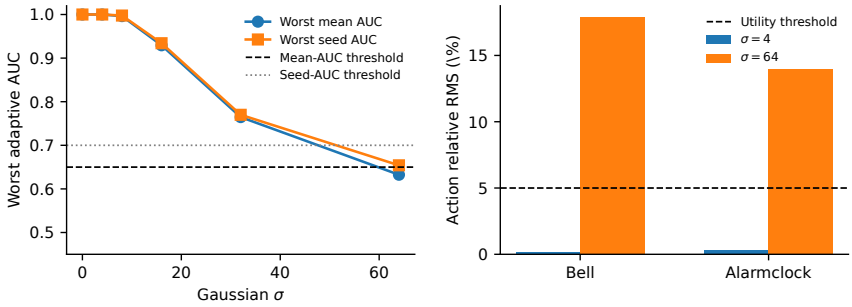


Figure 11: Only $\sigma = 64$ suppresses all retrained attacks, but at that level the induced action deviation is already too large for reliable control.

Table 9: Semantic projection restores closed-loop control to the Gaussian release while retaining low attack AUC. AUC maxima span four retrained attacks and four tasks. Each success column summarizes 120 RoboTwin episodes.

ϵ	worst mean AUC	worst seed AUC	cosine	raw DP success	projected DP success	Δ raw (pp)
0.1	.512	.536	.962	56.7	84.2	+1.7
0.3	.509	.521	.962	58.3	85.8	+3.3
1.0	.522	.541	.963	68.3	86.7	+4.2
3.0	.582	.603	.964	80.8	85.8	+3.3

For each episode, base OpenVLA processes the first frame with its recorded instruction, the same frame with a fixed within-task/role instruction permutation, and the midpoint frame with the recorded instruction. We retain the mean final-layer fused-prefix state, the mean of seven autoregressive action-decision states, and seven denormalized Bridge actions. Attack aggregates contain 20 episodes at fractions $\{0, .1, .2, .5\}$. Three split replicates generate 250 fit and 500 evaluation aggregates per fraction. All preprocessing is fit on the attacker role. Shuffled property labels and deterministic session-hash vectors are negative controls. This is an offline representation experiment. It neither executes OpenVLA on a physical robot nor evaluates real-world closed-loop success.

The fused-prefix logistic attack reaches task AUCs of 0.972, 0.961, and 0.998, with positive score direction throughout. Its shuffled-label interval contains chance, wrong-instruction mean AUC remains above 0.70, and the native signal exceeds the session-hash control by at least 0.10 AUC. By contrast, the action-decision state is inconsistent across tasks and predicted-action AUC is 0.532/0.505/0.569. The prespecified fused-prefix mixture regressor has task MAE 8.74/14.82/5.56 points, above the five-point recovery target. All 117 analysis cells and 1,755 stored metrics are independently recomputed from the raw score archive.

State-matched LiLa protocol. For each of four tasks, three mechanism seeds, 16 episodes from the randomized domain, and four temporal positions, we cache the same visual features, robot state, and initial flow noise before changing the condition. This yields 768 task–seed–query groups and 3,840 complete action chunks. A classifier fits only raw chunks from the first eight selected episodes per task and evaluates all five conditions on the disjoint final eight.

Private relative RMS is below 0.5% and cosine exceeds 0.99999 on every task. Wrong-task and zero exceed the larger private/public RMS by 0.165 (95% episode-cluster interval 0.096–0.221) and 0.139 (0.116–0.162), respectively, with positive direction in 4/4 tasks. Raw/private/public classifier accuracy is identically 96.875%, versus 60.938%/75.521% for wrong-task/zero. These comparisons isolate the first generated chunk and cannot be explained by later simulator divergence.

Table 10: Repeated release weakens the composed guarantee even though empirical attack AUC remains low. AUC entries are maxima over tasks and attack families. RDP composition uses $\delta = 10^{-5}$.

K	Mean AUC	Seed AUC	Gate	Basic ϵ	RDP ϵ
1	0.517	0.546	pass	1.0	1.01
2	0.513	0.540	pass	2.0	1.44
4	0.513	0.546	pass	4.0	2.07
8	0.528	0.553	pass	8.0	2.97

Table 11: Closed-loop success (%) under matched condition interventions. Each cell contains 30 episodes initialized from the same simulator seeds across conditions.

Condition	Bottle	Bell	Alarm	Switch	Pooled
Raw target	100.0	100.0	70.0	60.0	82.5
Private ϵ=1	96.7	100.0	80.0	70.0	86.7
Public anchor	96.7	100.0	76.7	70.0	85.8
Wrong task	0.0	0.0	20.0	23.3	10.8
Zero	0.0	100.0	43.3	26.7	42.5

I Secondary Properties and Interface Boundaries

Before feature inspection, we define initial-object workspace region and trajectory-duration strata from official metadata. The four-task analysis evaluates both labels wherever their definitions apply, including weak discrimination and poorly calibrated outcomes. These two properties are distinct from the object-instance and arm-side stress tests reported below.

The object-instance evaluation uses only demonstrations from the randomized domain and selects the lexicographically first of two upstream object assets as the positive group. No feature association determines this mapping. Adjust bottle and bell support both discrimination and fraction recovery. Alarmclock reaches AUC 0.998 but MAE 6.98 points. Table 13 reports every task.

Motion-intensity composition. Without opening visual features, we split each task's episodes from the randomized domain at the median mean consecutive 14-D joint-action displacement. The lower/upper counts are 244/244, 250/250, 250/250, and 242/242 for bottle, bell, alarmclock, and switch after excluding median ties. This property captures a deployment-relevant control-intensity mixture, not a human-sensitive label. At size 100, the multinomial attack distinguishes 0% from 10% low- intensity mixtures on all tasks (mean AUC 0.918–0.986, minimum-seed AUC 0.875–0.974), but its 9.46–11.25 point fraction MAE remains inaccurate.

We subsequently evaluate four continuous regressors trained only on shadow conditions. The best per-task MAE remains 8.75–9.09 points, so neither the original estimator nor this follow-up reaches the five-point recovery target (Table 14). We retain the strong hypothesis-testing result but do not claim calibrated mixture recovery for this property.

Alternate transition summaries. We also extract two non-deployed DINO summaries defined before evaluation. We use the layer--4 CLS difference and last-layer patch-token-mean difference. Both support discrimination and fraction recovery on all four tasks (Table 15). This controls one representation-design choice inside the same visual encoder. It does not constitute cross-WAM validation.

Independent image encoders. Official CLIP ViT-B/32 and SigLIP base patch16-224 goal-minus-start embeddings were extracted for all four tasks under shared preprocessing. At aggregate size 100, the logistic attack obtains mean/minimum-seed AUC 1.000/1.000 for both encoders on every task.

Table 12: Where collection information appears in OpenVLA on BridgeData V2. The fused prefix reaches 0.977/0.961 mean/minimum-task AUC, while shuffled labels, session identity, and later action observables constrain the finding. "Many" denotes the many-skills task.

Observable	Drawer	Many	Stack	Mean	Min.
Fused prefix	0.972	0.961	0.998	0.977	0.961
Action-decision state	0.753	0.503	0.862	0.706	0.503
Predicted action	0.532	0.505	0.569	0.535	0.505
Shuffled-label fused	0.499	0.500	0.499	0.499	0.499
Session-hash control	0.409	0.507	0.505	0.474	0.409

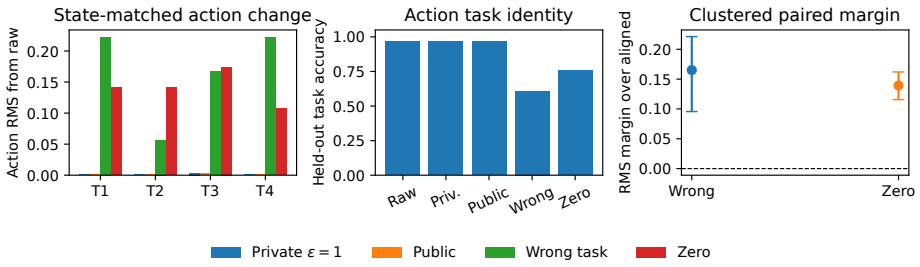


Figure 12: Condition effects before trajectories diverge. Left. Action change relative to the raw condition. Center. Task identity in the first action chunk. Right. Episode-clustered contrasts against wrong-task and zero controls. Observations and flow noise are shared within every comparison.

CLIP fraction MAE ranges from 0.04 to 1.16 points. SigLIP fraction MAE ranges from 0.04 to 1.86 points. Both encoders therefore support discrimination and accurate recovery across tasks. This is a cross-encoder representation result, not evidence that either encoder is a deployed WAM condition interface.

OpenCLAP and DW05. Using the official OpenCLAP checkpoint, the continuous quantized latent-action summary supports both objectives on all three evaluated tasks. Logistic AUC spans 0.976–1.000 and fraction MAE spans 1.73–2.83 points. Its token histogram succeeds on only two tasks and serves as a contrasting representation. With the official DW05 checkpoint and disclosed two-camera adapter, normalized and denormalized action-output aggregates at size 20 yield logistic AUC 0.978–0.996 on every task. Their 6.06–9.62 point MAE remains above the five-point recovery target. Thus cross-model discrimination is strong, calibrated mixture recovery is representation-dependent, and neither experiment is presented as a native released task-condition interface.

Zero-WAM collection-adapter boundary. Using 180 official HumanGen latents at one pinned revision, we form an external shape-matched mean-latent release and keep target text fixed. Three source-disjoint splits support four tasks and K in $\{2, 4, 8, 16\}$. At $K = 16$, the histogram estimator does not jointly achieve high AUC, low-FPR detection, and accurate recovery on any task, although logistic endpoint discrimination remains strong. A subsequent recovery analysis fits four fraction regressors at K=$16/32/64$ and selects only by calibration MAE. By K=64, minimum-seed AUC is 0.917–1.000 and TPR@1% FPR is 0.462–1.000, yet worst-seed MAE remains 18.0–27.8 points. Thus aggregation amplifies binary target-domain presence but does not identify mixture proportion. These results concern adapter representations. The native forward conditions on a single demonstration video rather than a released multi-demonstration aggregate.

We examine six additional interfaces comprising DreamZero, OpenDW, OpenCLAP, A2World, Zero-WAM, and LiLa's second condition bank. Three support official forwards and two support complete four-task representation studies. Zero-WAM supplies a second native demonstration-conditioned control interface, but it conditions on one video rather than a released collection aggregate. None

Table 13: Object-instance composition inferred from size-80 conditions built without replacement.

Task	Variant counts	Mean AUC	Min AUC	TPR@1%	MAE (pp)
Adjust bottle	297/192	1.000	1.000	1.000	0.75
Click bell	239/261	0.999	0.999	0.985	1.92
Click alarmclock	281/219	0.998	0.996	0.958	6.98

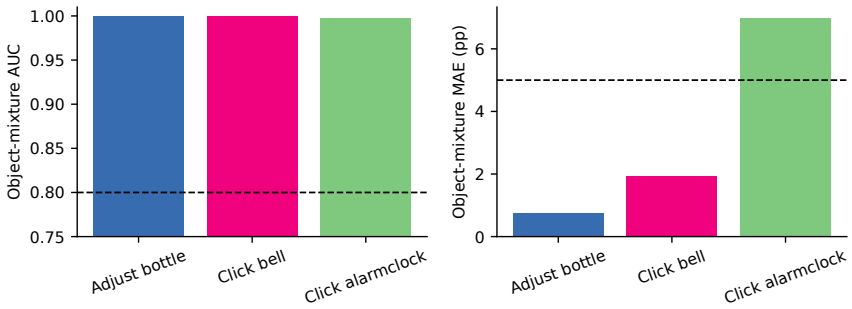


Figure 13: Object-instance composition is both distinguishable and recoverable for bottle and bell. Alarmclock remains distinguishable but poorly calibrated.

therefore shares LiLa's privacy unit. This distinction limits the native cross-WAM claim while leaving the constructed OpenCLAP and DW05 representation results intact.

For the arm-side experiment, only episodes from the randomized domain are used, so the primary collection-domain label cannot confound the property. Episode HDF5 names match official scene metadata without omissions. Table 19 shows strong discrimination, while fraction error falls below five points on only one task.

J Qualitative Rollout Evidence

Rendering protocol. For each task and evaluation domain, the expert sequence is the start, temporal midpoint, and final frame of the lowest-index valid trajectory. The paired policy pages use the first prevalidated simulator seed for each task and show private, public-anchor, wrong-task, and zero conditions. Their start frames are pixel-identical across conditions, and every selected outcome is displayed without success-based filtering. These renderings provide qualitative context. all performance claims use the complete rollout matrices.

Table 14: Motion-intensity composition at aggregate size 100. Continuous regressors improve fraction recovery only modestly and are selected using shadow conditions alone.

Task	Logistic AUC	Min AUC	MAE (pp)	Best recovery MAE	Joint pass
Adjust bottle	0.918	0.875	11.25	9.09	no
Click bell	0.979	0.968	9.46	8.80	no
Click alarmclock	0.986	0.974	10.09	8.75	no
Turn switch	0.958	0.942	10.89	8.79	no

Table 15: Collection leakage under two alternate DINO transition summaries, with each aggregate sampled without replacement.

Task	Representation	Mean AUC	Min AUC	MAE (pp)	Pass
Adjust bottle	CLS, layer -4	1.000	1.000	0.09	yes
Adjust bottle	Last-layer patch mean	1.000	1.000	0.02	yes
Click bell	CLS, layer -4	1.000	1.000	0.16	yes
Click bell	Last-layer patch mean	1.000	1.000	0.15	yes
Click alarmclock	CLS, layer -4	1.000	1.000	1.32	yes
Click alarmclock	Last-layer patch mean	1.000	1.000	0.99	yes
Turn switch	CLS, layer -4	1.000	1.000	0.03	yes
Turn switch	Last-layer patch mean	1.000	1.000	0.04	yes

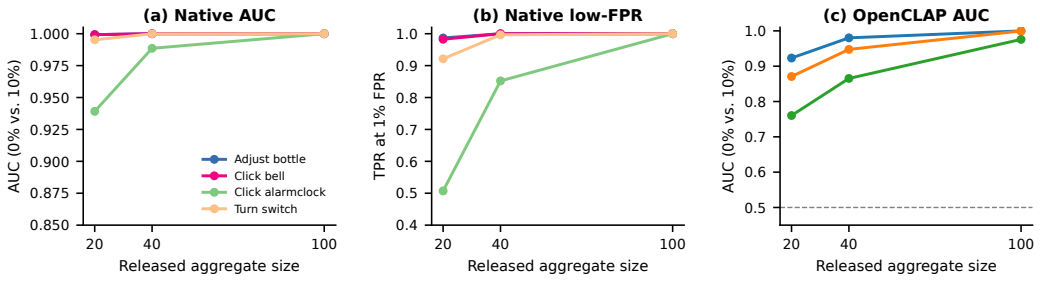


Figure 14: Larger aggregates sharpen collection inference under episode-disjoint evaluation. **(a)** LiLa AUC and **(b)** low-FPR discrimination both increase with size. **(c)** OpenCLAP exhibits the same trend. The dashed line denotes chance.

Table 16. Collection inference from OpenCLAP and DW05 representations. MAE is fraction error in percentage points.

Model	Task	AUC	Min AUC	MAE (pp)	TPR@1%
OpenCLAP	Adjust bottle	1.000	1.000	1.73	99.8
OpenCLAP	Click bell	1.000	0.999	2.83	98.7
OpenCLAP	Click alarmclock	0.976	0.955	2.72	63.0
DW05	Adjust bottle	0.996	0.995	9.62	94.1
DW05	Click bell	0.982	0.957	6.71	76.7
DW05	Click alarmclock	0.979	0.972	6.07	76.7

Table 17. Collection inference from independent CLIP and SigLIP visual representations.

Encoder	Task	AUC	Min AUC	MAE (pp)
CLIP	Adjust bottle	1.000	1.000	0.04
CLIP	Click bell	1.000	1.000	0.26
CLIP	Click alarmclock	1.000	1.000	1.16
SigLIP	Adjust bottle	1.000	1.000	0.04
SigLIP	Click bell	1.000	1.000	0.40
SigLIP	Click alarmclock	1.000	1.000	1.86

Table 18: Collection inference from Zero-WAM adapter representations. HGB columns summarize the $K = 16$ histogram estimator. Logistic gives an independent $K = 16$ endpoint comparison, and recovery MAE reports the strongest continuous regressor selected on shadow conditions.

Task	HGB AUC	HGB TPR	HGB MAE	Logit. AUC	Reg. MAE_{16}	Reg. MAE_{32}	Reg. MAE_{64}
Scale	0.730	0.062	26.1	0.825	23.7	20.9	21.1
Stapler	0.797	0.000	21.4	0.912	20.8	18.9	21.4
Stamp	0.783	0.000	22.4	0.967	18.9	21.3	18.0
Cup	0.741	0.103	25.1	0.981	27.8	27.6	27.8

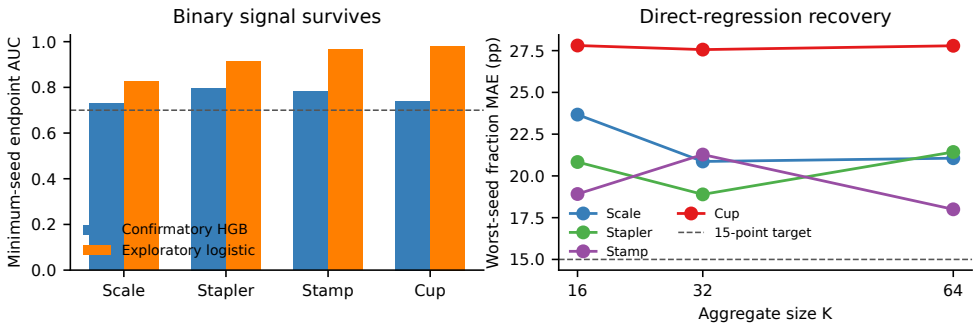


Figure 15: Zero-WAM adapter representations reveal binary target-domain presence more clearly as the collection grows, but fraction estimates remain above 15 percentage points of error through $K = 64$.

Table 19: Arm-side composition inferred by logistic regression from conditions from the randomized domain.

Task	Left/right	Mean AUC	Min AUC	MAE (pp)	TPR@1%	FPR@95%
Adjust bottle	242/247	1.000	1.000	1.82	1.000	0.000
Click bell	235/265	0.984	0.963	6.40	0.766	0.091
Click alarmclock	244/256	0.943	0.925	8.93	0.505	0.265

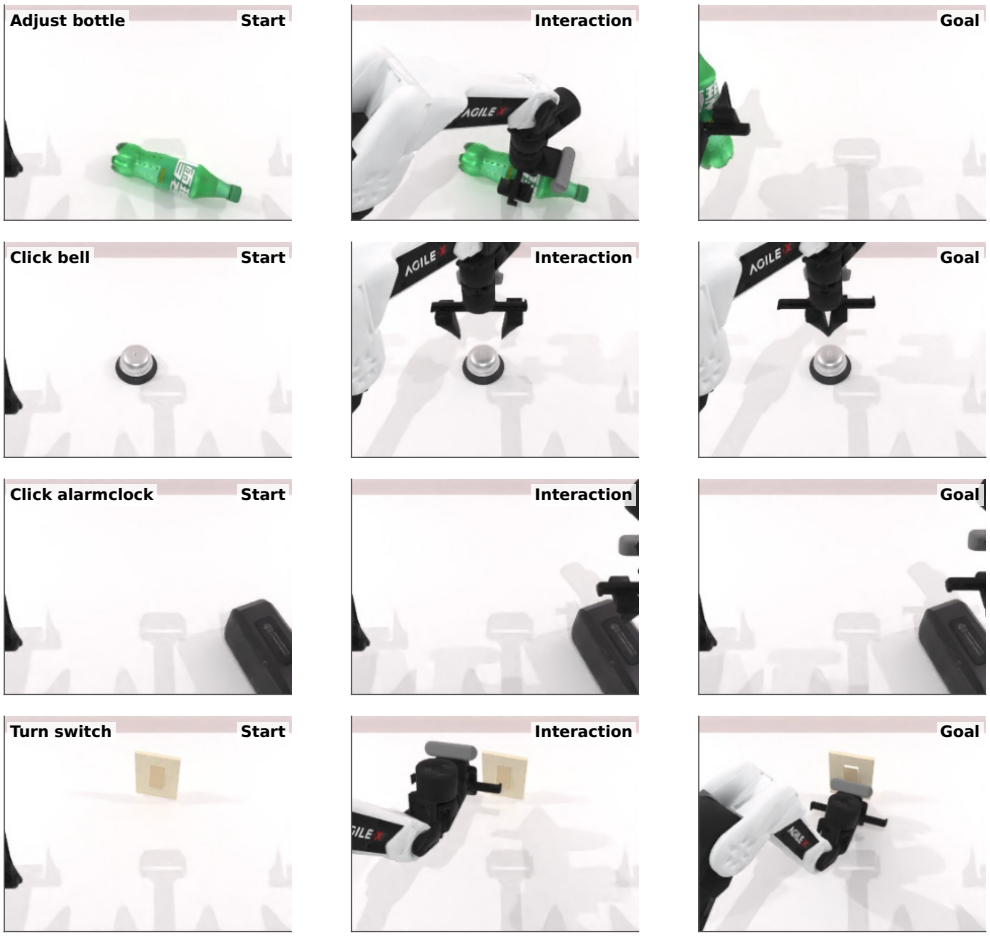


Figure 16: Nominal-domain RoboTwin demonstrations used to form task conditions. Each row follows the lowest-index valid expert episode from start through interaction to completion, illustrating task semantics rather than estimating success.



Figure 17: Randomized-domain demonstrations in the same start–interaction–goal layout as Fig. 16. Background, texture, lighting, and distractors change substantially while task semantics remain constant. The quantitative analysis uses the full episode set.

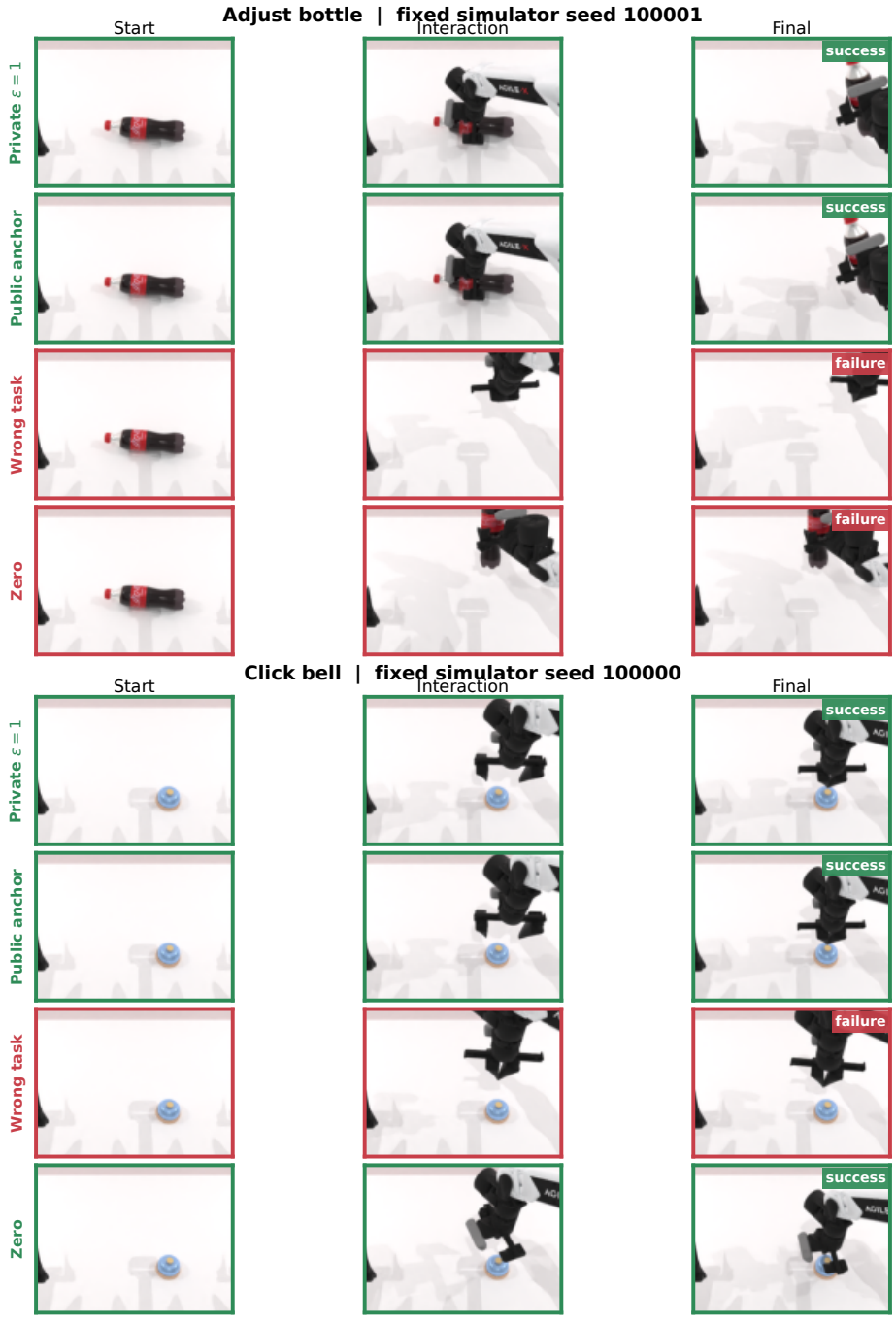


Figure 18: Matched policy sequences for adjust bottle and click bell. Rows vary only the condition. Columns show the shared start, an intermediate observation, and the final frame. Green and red borders indicate the recorded outcome. For each task, we show the lowest-index valid simulator seed irrespective of outcome.

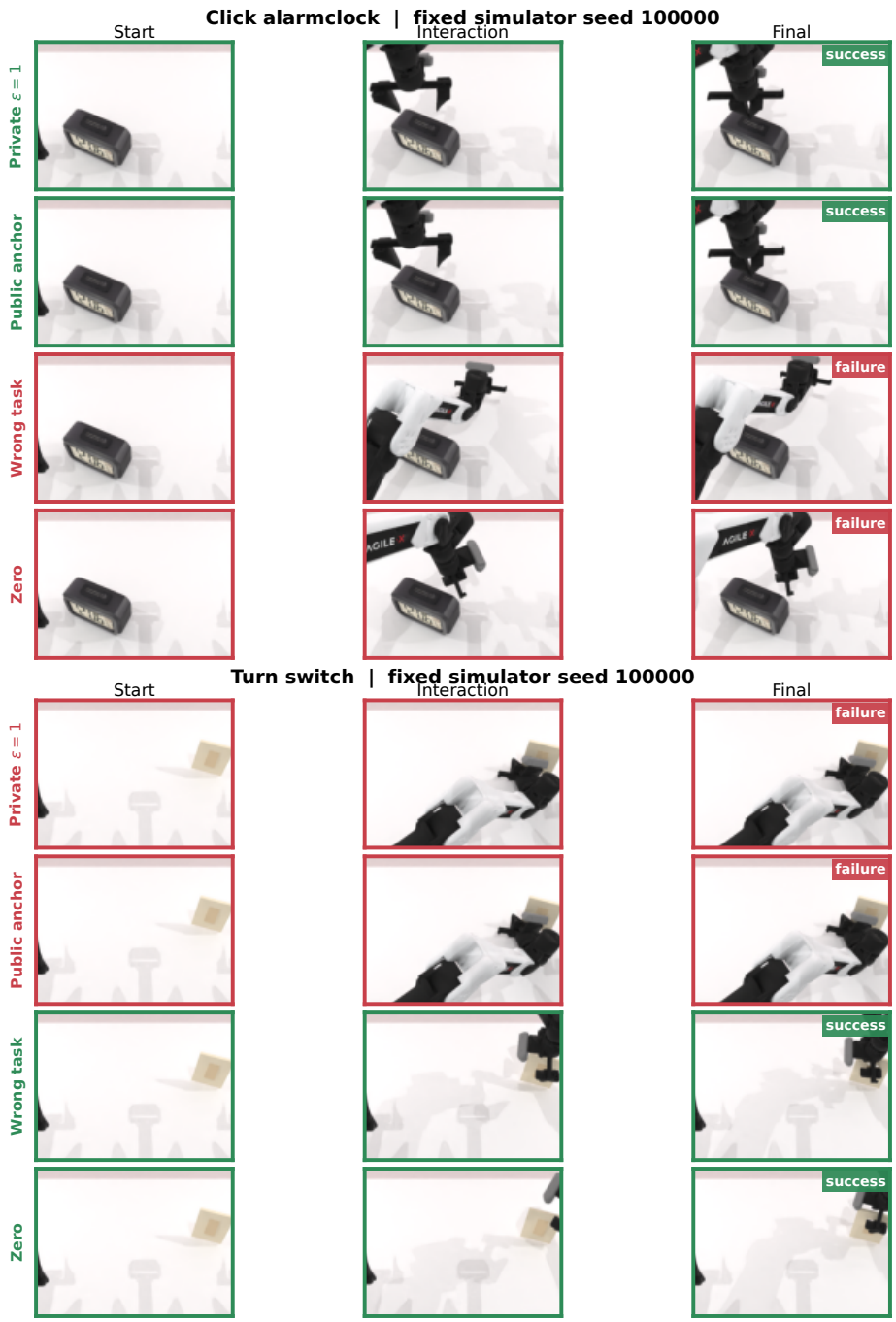


Figure 19: Matched policy sequences for click alarmclock and turn switch, continuing Fig. 18. The contrasting outcomes illustrate condition-dependent behavior. Success rates are computed from the complete rollout matrices rather than these examples.